DOCUMENT RESUME

ED 395 027 TM 025 045

AUTHOR Stocking, Martha L.
TITLE Empirical Estimation Errors in Item Response Theory as a Function of Test Properties.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-5
PUB DATE Feb 89
NOTE 79p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Item Response Theory; *Maximum Likelihood Statistics; Models; Test Construction; *Test Items
IDENTIFIERS BILOG Computer Program; Calibration; Empirical Research; *LOGIST Estimation Procedures; Marginal Maximum Likelihood Statistics

ABSTRACT

The success of applications of item response theory (IRT) depends upon the properties of the estimates of model parameters. Many theoretical properties of these estimates have been extensively studied. However, the properties of estimates obtained empirically from real data depend not only on the theoretical results, but also on the data and the estimation procedures used to obtain them. In this paper, the properties of estimates obtained from a commonly-used implementation of the joint maximum likelihood approach (LOGIST) are examined extensively and shown to be, in part, functions of the properties of the test or item set being calibrated. A small study is also made of the properties of estimates obtained from a commonly-used implementation of the marginal maximum likelihood approach (BILOG). Recommendations are made for the improvement of both procedures. (Contains 3 tables, 23 figures, and 29 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 395 027

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

EMPIRICAL ESTIMATION ERRORS IN ITEM RESPONSE THEORY AS A FUNCTION OF TEST PROPERTIES

Martha L. Stocking

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
February 1989

TM025045

Empirical Estimation Errors in Item Response Theory as a Function of Test Properties*

Martha L. Stocking

February 1989

*This work was supported by ETS through the Program Research Planning council.

Copyright © 1989. Educational Testing Service. All rights reserved.

Abstract

The success of applications of Item Response Theory (IRT) depends upon the properties of the estimates of model parameters. Many theoretical properities of these estimates have been extensively studied. However, the properties of estimates obtained empirically from real data depend not only on the theoretical results, but also on the data and the estimation procedures used to obtain them. In this paper, the properites of estimates obtained from a commonly-used implementation of the joint maximum likelihood approach (LOGIST) are examined extensively and shown to be, in part, functions of the properties of the test or item set being calibrated. A small study is also made of the properties of estimates obtained from a commonly-used implementation of the marginal maximum likelihood approach (BILOG). Recommendations are made for the improvement of both procedures.

Key words: IRT empirical estimation errors
IRT empirical bias
test properties and IRT estimation
LOGIST
BILOG

Empirical Estimation Errors As a Function of Test Properties

Introduction

The theoretical advantages of Item Response Theory (IRT) psychometric models over classical test theory are by now well known and appreciated in the educational and psychological measurement communities. When model assumptions are satisfied, true item parameters do not change even when the items are subsets of larger item sets or when parameters are estimated from examinees with different true abilities sampled from the same population. Likewise, true abilities do not change, even when estimated from different sets of items (Lord, 1980). This 'invariance' property of true item and person parameters provides the foundation for many important applications of IRT.

Some of the more novel applications include new test designs such as computerized adaptive testing (see Lord, 1980, Chapter 10; Weiss, 1975; Holland et al., 1988; Stocking, 1987), mastery test design (Lord, 1980, Chapter 11; Lewis & Sheehan, 1988), new test development paradigms (Birnbaum, 1968; Lord, 1980; van der Linden & Boekkooi-Timminga, 1987), and new equating and pre-equating methodologies (see Lord, 1980, Chapter 13; Cook, Petersen, & Stocking, 1983; Eignor & Stocking, 1986; Stocking & Eignor, 1986). The extent to which these and other applications of IRT are

successful in practice depends not on the properties of the true item and person parameters, but rather on the properties of estimates of the item and person parameters.

The purpose of the present paper is to explore, through a series of Monte Carlo simulation studies, the degree to which and some circumstances under which estimates of item and person parameters fail to achieve the invariance properties of true item and person parameters, and some practical consequences. The heart of the discussion concentrates on a situation commonly encountered in practice: the estimation of parameters of sets of items (for convenience these item sets will be referred to as tests) and the simultaneous or subsequent use of these estimates to obtain estimates of examinee ability. The context is the 3-parameter logistic (Birnbaum) item response model.

The major focus of this paper is the joint maximum likelihood estimation procedures for obtaining parameter estimates as incorporated in the widely used computer program, LOGIST (Wingersky, Barton, & Lord, 1982). A small investigation is made of the more recently developed marginal maximum likelihood approach incorporated in the computer program BILOG (Mislevy & Bock, 1983). These two programs incorporate very different approaches to the problem of obtaining estimates of item and

person parameters (Mislevy & Stocking, 1987), but suffer some of the same deficiencies.

The Three-Parameter Logistic Item Response Model

Central to item response theory is a mathematical expression for the probability that an examinee with ability θ (theta) will correctly respond to a particular test item. Under the three-parameter logistic (3PL) model for test items that are scored either right or wrong (Birnbaum, 1968), this probability, denoted by $P(\theta)$ has the following form:

$$P(\theta) = c + \frac{(1 - c)}{1 + e^{-1.7a(\theta-b)}} \qquad (1)$$

where a, b, and c are parameters characterizing the item and e is the mathematical constant. The three item parameters have specific interpretations. The c parameter is the probability that an examinee completely lacking in ability will answer the item correctly, and is frequently called the guessing parameter. The b parameter determines the location of the curve along the ability scale. This parameter characterizes the difficulty of an item, in that if a and c are held constant, higher values of b imply lower probabilities of correct response from all examinees. The logistic curve has its point of inflection at $\theta = b$. The a

BEST COPY AVAILABLE

parameter is proportional to the slope of the curve at this inflection point. This a parameter characterizes the discrimination of the item, in that probabilities of correct response to items with high a values are more sensitive to changes in ability in the neighborhood of the item difficulty.

An Important Distinction

For statisticians, the theoretical properties of statistical estimators are of primary importance. For practitioners, who deal with the results of implementations of statistical procedures, these theoretical properties are less important than the empirical estimation errors and biases actually obtained in practice. It is the latter that are the focus of this paper.

Joint maximum likelihood (JML) and marginal maximum likelihood (MML) are statistical procedures for obtaining parameter estimates. As such, they have certain theoretical properties that have been studied extensively. Lord (1983a) derives the theoretical bias and standard errors of maximum likelihood ability estimates derived from known item parameters. Lord (1983b) derives the theoretical bias and standard errors of maximum likelihood item parameter estimates derived from known abilities. Lord and Wingersky (1985) develop standard errors of item and ability parameter estimates when all parameters are

estimated jointly by maximum likelihood. Lord (1986) considers the theoretical bias and standard errors of maximum likelihood, marginal maximum likelihood, and Bayesian estimation procedures. Davey and Levine (1988) compare the theoretical standard errors for item parameter estimates assuming known abilities with those obtained when abilities are estimated simultaneously.

The empirical properties of either JML or MML estimators differ from the theoretical properties for many reasons. In the present instance, LOGIST does not typically produce JML estimates of item and person parameters; BILOG does not typically produce MML estimates of item parameters and Bayesian estimates of abilities (Mislevy & Stocking, 1987). Rather, the procedures produce approximations that are considered to be 'good enough'. Even if LOGIST is required to produce JML estimates and BILOG required to produce MML estimates, both LOGIST and BILOG results depend not only upon the information contained in the response data, but also upon information supplied by the researcher as may be required to produce reasonable and efficient solutions (e.g., starting values, boundaries, prior distributions, etc.). Thus the empirical properties of estimates produced by either program can be expected to differ across tests and across samples of examinees. The simulation studies described in subsequent

sections indicate that some of the empirical properties of the two estimation procedures are less than optimal and, possibly, correctable.

The Correlations Among Errors of Estimation for Item Parameters

Wingersky and Lord (1984) study the theoretical estimation errors for item and person parameters when estimated by JML. They note that:

1) The correlation between errors made when estimating item discrimination and item difficulty "is moderately or strongly positive for easy items and moderately or strongly negative for difficult items."

2) The correlation between errors made when estimating item difficulty and the guessing parameter, as item discrimination decreases, "becomes strongly positive except for difficult items where the guessing parameter is well determined."

3) The correlation between errors made when estimating item discrimination and the guessing parameter "for difficult items ... is positive and sometimes high; for easy items the correlation is negative."

These theoretical correlations between estimation errors are not a specific or unique consequence of the JML estimation procedure, nor any other estimation procedure. Rather, they are a

consequence of the fact that the mathematical model postulated for the item response function has a constrained shape (in the case of the 3PL, this shape is symmetric), and that estimation procedures have available data from only a limited region to estimate parameters that describe this shape over the entire range of ability. To the extent that the limited item response data fits the postulated model, a particular missestimation of one item parameter causes an automatic and predictable adjustment in the estimates of the other parameters in order to fit the available data as well as possible.

To illustrate this point, Figure 1 shows the genesis of the correlation between errors of sample estimates of item discrimination and item difficulty. In each subfigure, the solid curve is the true item response function that the data are assumed to follow. For the purpose of this illustration, we assume that the distribution of ability from which examinees are drawn is bell-shaped; most of the observed data fall in the middle region of each panel. The top row shows an item that is too hard for most examinees ($a = 1.0$, $b = 2.0$, $c = .2$); the bottom row shows an item that is too easy for most examinees ($a = 1$, $b = -2$, $c = .2$).

Suppose, for a hard item, an item discrimination is underestimated (top left panel, $\hat{a} = .5$). If the difficulty of the

item is then estimated to be equal to the true difficulty ($\hat{b}$ = 2.0), the resulting item response function (shown with long dashes) does not lie very close to the true item response function in the middle regions where most of the data lie. If the difficulty of the item is estimated to be less than the true difficulty ($\hat{b}$ = 1.0), the resulting item response function (shown with short dashes) lies even further away from the true item response function in this middle region. Only by estimating the difficulty of the item to be larger than the true difficulty ($\hat{b}$ = 3.0) can a close fit to the true item response function be obtained in the middle regions. An analagous argument holds when, for a hard item, the item discrimination is overestimated (top right panel, $\hat{a}$ = 1.5). The estimated difficulty will then be underestimated ($\hat{b}$ = 1.5) in order to fit the preponderance of the data. These two panels illustrate the Wingersky and Lord negative correlation between errors of sample estimates of item discrimination and item difficulty for hard items.

A similar logic leads to the positive correlation for an easy item in the two panels of the bottom row. When the item discrimination is underestimated ($\hat{a}$ = .5), underestimating the item difficulty ($\hat{b}$ = -3.0) moves the estimated curve closest to the true curve. When the item discrimination is overestimated

($\hat{a} = 1.5$), overestimating the item difficulty ($\hat{b} = 1.5$) has the same effect.

Insert Figure 1 about here

These correlations or errors of estimates may be, and probably are, important in most applications of IRT. They assume additional importance if estimated item parameters are used as if they had the properties of true item parameters in contexts different from the original calibration of such items. This is so because these correlations among errors of estimation induce biases in the estimated item parameters which may propagate in unknown ways in these new contexts.

The General Simulation Design

Although the details of some of the simulations differ, and will be described in subsequent sections, most follow a general pattern summarized here. In all but one of the simulations reported, samples of simulated examinees (simulees) remain constant, while the properties of the collection of items to be calibrated vary. Typically the sample of simulees is drawn from a rectangular distribution of true ability. Such samples should be better than those drawn from typical (bell-shaped) distributions

of true ability; Wingersky and Lord (1984) show that such samples produce smaller errors in both item and ability parameter estimation when the JML procedure is used. A set of true item parameters is then used to define a test, and response data are generated for each simulee using the true ability, true item parameters, and the 3PL model of equation (1). A response is assigned to an item/simulee encounter by comparing the modeled probability with a random draw from a uniform distribution on the unit inverval. If the modeled probability is greater than the random number, a correct response is assigned; otherwise an incorrect response is assigned.

Item parameter estimates are obtained (the test is calibrated) using the JML procedure and LOGIST in all but one case, in which the MML procedure and BILOG are used for comparison. Since the calibration results are reported on a metric specific to a particular calibration, the results are transformed to the metric of the true parameters using the Stocking and Lord (1983) transformation method. This method finds as parameters of the linear transformation between two different metrics those values that minimize the difference between the two test characteristic curves.

Transformed item parameter estimates are then compared with their true values by plotting the residual (estimated value minus true value) against the true value. If there is no estimation error or bias in the item parameter estimates, these residuals lie on a horizontal line at zero. If the estimates contain error, but no bias, these residuals form a cloud of points that is evenly bisected by the horizontal line at zero. If the estimates also contain bias, the point cloud may take on different shapes.

Ability estimates from a typical LOGIST calibration are not computed from the final item parameter estimates. Using the default procedures of the LOGIST program the final program step improves the item parameter estimates while fixing the ability estimates. To obtain ability estimates based strictly upon the final item parameter estimates, a criterion or cross-validation sample of simulees is independently drawn from the original true distribution of simulee ability, and ability estimates for this criterion sample are obtained assuming that the transformed item parameter estimates are in fact the true parameters. These estimated abilities are compared with their true values by first computing the median of the estimated abilities for small intervals of true abilities. Lower and upper limits of a nonparametric two-tailed 5% confidence band around the median are

also computed using the method of David (1981). The residual of this median (median estimated ability minus true ability) and also of the confidence band are then plotted against the true ability. If the estimated abilities contain no estimation error or bias, the residuals and confidence bands lie on the horizontal line at zero. If the estimated abilities contain estimation error, but no bias, the residuals of the medians fluctuate slightly around the horizontal line at zero, and the confidence bands include zero. If the estimated abilities are biased, there are portions of the range of true ability in which the residuals of the medians differ from zero and for which the confidence intervals do not contain zero.

Four Extreme Tests

We want to show that the relationship between true and estimated parameters is determined by the peculiarities of the test. To illustrate this point the four tests for this simulation were chosen to yield strikingly different relationships. Each test consisted of 100 5-choice items. The true c's for each item in each test were set equal to .15. This value was chosen based on the observation that in practice c's are usually estimated as smaller than the probability of a correct answer based on random guessing. The true b's for each test were chosen randomly from a

rectangular distribution with a range of -2.5 to +2.5. The true a's were always chosen to be within the range of .5 to 2.5, but with different correlations with the true b's for each test. These correlations varied as follows:

1) For test C1, the population correlation between true a and true b was +.8, and the sample correlation was +.76 .

2) For test C2, the population correlation was -.8, and the sample correlation was -.76.

3) For test C3, items with difficulty less than zero (moderate to easy items) had a population correlation with item discrimination of -.8. Items with difficulty greater than zero (moderate to difficult items) had a population correlation with item discrimination of +.8. The overall sample correlation attained was .08.

4) For test C4, moderate to easy items had a population correlation with discrimination of +.8, and moderate to difficult items had a population correlation with discrimination of -.8, with an overall sample correlation of .00.

General Results

Table 1 summarizes the true item parameters of these tests. Of the four tests, test C1 is the most realistic. For real tests, there is typically a positive correlation between item

discrimination and item difficulty (see for example, Lord (1975a)), although certainly not as strong as this. However, the situations represented by C2, C3, and C4 may not be so unusual if the items to be calibrated come from many different tests or subtests calibrated simultaneously.

Insert Table 1 about here

A calibration sample of N = 3000 simulees was chosen from a rectangular distribution of true ability from -2.5 to +2.5. The population mean and standard deviation are 0.0 and 1.44; while the sample mean and standard deviation were -.01 and 1.46. A separate criterion or cross-validation sample of another N = 3000 simulees was chosen from the same distribution, with sample mean and standard deviation of -.03 and 1.45. Item response data for each test were generated for each simulee in each sample. For the calibration sample, these data are used to obtain item parameter and ability estimates from LOGIST. For the cross-validation sample, these data are used to obtain only ability estimates.

Lord (1983a) derives the theoretical bias and standard error of maximum likelihood ability estimates, assuming that the item parameters are known (true) values. The theoretical bias,

although small, suggests that high true abilities are slightly overestimated, and low true abilities are slightly underestimated. Figure 2 shows the residual median plots, described earlier, for the cross-validation sample when the ability estimates have been computed from the generated response data and the true item parameters for each test. The theoretical bias predicted by Lord appears insignificant in these plots.

Insert Figure 2 about here

Each test was calibrated, using the generated response data for the calibration sample, and transformed to the metric of the true parameters. The final transformed item parameter estimates, along with the generated data for the cross-validation sample were then used to computed new estimated abilities. Figure 3 shows the residual median plots for these estimated abilities. It is clear that abilities estimated from estimated item parameters (Figure 3) are very different from abilities estimated from true item parameters (Figure 2).

Insert Figure 3 about here

It is well to remember when examining these Figures that the calibration sample of simulees was the same for all four tests, and the cross-validation sample of simulees is also the same for all four tests. The precision of the estimates, as judged by the width of the confidence bands, appears about the same using either true or estimated item parameters to estimate ability, and across all four tests. While there is no apparent bias in the ability estimates when obtained from true item parameters, the bias is significant when ability estimates are obtained from estimated item parameters. And, in spite of the fact that the calibration and cross-validation samples are the same for each setting, the bias differs by test.

Figures 4, 5, 6, and 7 show the residuals for the estimated item parameters for each of the four tests. Each Figure has three panels -- the top panel shows the residuals for the estimated a's, the middle panel shows the residuals for the estimated b's, and the bottom panel shows the residuals for the estimated c's. For the residuals for the item difficulties, items with overestimated discriminations are plotted with a plus for a plotting symbol; items with underestimated discriminations are plotted with a hexagon.

Insert Figures 4, 5, 6, and 7 about here

In each of the four Figures, the overall patterns of bias for the b's mirrors, as it must, that of the bias in the estimated abilities for the criterion sample shown in Figure 3. In addition, each of the figures shows the negative correlation between sample estimates of a and b for hard items, and the positive correlation for easy items.

For test C1, with a positive correlation of .8 between the true a and true b, high b's are overestimated and have underestimated a's; low b's are overestimated and have overestimated a's. Very high a's are underestimated, and low a's tend to be overestimated. For test C2, with a negative correlation of -.8 between true a and true b, high b's are underestimated with overestimated a's; low b's are underestimated with underestimated a's. The estimated a's are more widely scattered than for test C1, but also show high a's as underestimated and low a's as overestimated. Test C3 has a negative correlation between true a and b for moderate to easy items and a positive correlation for moderate to difficult items. Difficult items are overestimated with underestimated a's; easy

items are underestimated with underestimated a's. More a's are underestimated, and almost all moderate to high a's are underestimated; most of the lower a's are overestimated. For test C4, with a positive correlation between true a's and b's for moderate to easy items and a negative correlation for moderate to difficult items, difficult items have underestimated b's and overestimated a's; easy items have overestimated b's and overestimated a's. Most of the a's are overestimated.

Test C2 and LOGIST

The dependence of the bias in item parameter estimates and ability estimates upon the properties of the test is demonstrated in Figures 2 through 7. If typical samples with bell-shaped distributions of ability had been used, one might conclude that this dependence occurs from lack of available data with which to estimate precisely the more extreme (high and low) difficulties or abilities. But this is not the case here since the calibration sample is rectangular. Except for possibly the most extreme difficulties and abilities, approximately the same amount of information is available in the response data for estimating the parameters of all items.

An hypothesis was developed and explored to explain the dependence of the bias on test properties based on the

characteristics of the calibration procedure used, LOGIST. As a default, LOGIST imposes a particular structure on the alternating stages of ability and item parameter estimation. A partial rationale for this structure is given in Mislevy and Stocking (1987). This structure consists of series of four steps within which restrictions are imposed upon what parameters are actually being estimated. In Step 1, a's and c's are held fixed at their starting values, and only abilities and difficulties are estimated. In Step 2, abilities are fixed at current values, and only the item parameters are estimated. Step 3 is like Step 1, and Step 4 is like Step 2. The effect of this structure is to prevent the procedure from being run to complete convergence, while providing reasonable estimates for item parameters and abilities in a feasible amount of time.

Consider test C2 with a correlation of -.8 between true a and true b. In Step 1, abilities and difficulties are estimated, while discriminations and guessing parameters are held fixed at their starting values. The starting values for the c's are typically all equal to a default value set by the program to be slightly lower than the probability of a correct response by chance. The starting values for the discriminations are typically all equal to a common middle value specified by the researcher and

based on previous experience with similar data. To mirror this situation, in the LOGIST runs for tests C1 through C4, the initial a parameter was set to the mean of the true a parameters for a run.

For test C2, this starting value of a will be lower than the true a's for easy items, and higher than the true a's for the hard items. This can, and probably does, induce the correlation in Step 1 between the errors of estimated difficulties (different for each item) and this common a value. If the bias is present in the estimated difficulties, it will also be present for the estimated abilities since biased difficulties and equally biased abilities yield the same modeled probabilities for the observable responses that LOGIST is attempting to fit. In Step 2, these estimated abilities are fixed (at their biased values) and a, b, and c are estimated. Because the biased abilities are fixed, the biased estimated b's cannot change much, causing the individual a's and c's estimated in Step 2 to accommodate to the bias. Step 3 now fixes the estimated a's and c's and reestimates the abilities and difficulties. However, the a's and c's are already biased so the abilities and difficulties remain so. Step 4, fixing the biased estimated abilities and estimating the item parameters for the last time cannot undo the bias.

The implication of this scenario is that the bias is introduced by the initial common value for the item discrimination with which the LOGIST procedure begins. To test this hypothesis, the calibration of test C2 was repeated with better starting values for the item discriminations. Instead of a common value, the starting value for each item discrimination was set to the value of the true discrimination. It bears emphasis to note that the true discriminations were used only as starting values; they were not fixed throughout the LOGIST run.

The residuals for the item parameter estimates are shown in Figure 8. When compared to the previous results for this test shown in Figure 5, the correlation between errors of estimates for a and b are barely visible and only present for the most extreme difficulties. The discriminations are slightly overestimated throughout. The effects on estimated abilities for the calibration sample (not the cross-validation sample) are shown in the top panel of Figure 9. Results for the calibration sample can be expected to be more extreme than those for the cross-validation sample since they are not computed from the final item parameter estimates.

Insert Figures 8 and 9 about here

While it is clear that better starting values for the item discriminations removes a large part of the bias seen in Figure 5 and Figure 3, it is somewhat startling to find that changing starting values for item discriminations has such a large effect on the standard LOGIST procedure. Although the default four-step procedure was known to prevent complete convergence, it was not known to be especially sensitive to starting values. These results give rise to the speculation that <u>if LOGIST were run to complete convergence the biases seen in the more typical LOGIST run with poor starting values for item discriminations might also disappear</u>.

To test this hypothesis, the calibration of test C2 was repeated twice more. In each of these runs, the four-step procedure was bypassed and LOGIST was allowed to run to complete convergence. In the first of these two runs, the initial a parameters were set at the common value, as before, and in the second run the initial a parameters were set at their true values. The results, in terms of the estimated abilities for the calibration sample, are shown in the middle and bottom panels of

Figure 9. Both of these runs produce less satisfactory results than the four-step procedure with true a's as starting values (top panel of Figure 9). Running LOGIST to complete convergence allows too much movement away from the good starting values.

It is reassuring to note that when running LOGIST to complete convergence, the starting values of the item discriminations have no impact on the final results (compare the middle and bottom panels). It is not reassuring to note that, although substantially reduced, the same bias as noted originally in Figure 3 for the residual medians is still present when LOGIST is run to complete convergence.

The item parameter estimates from these two complete convergence runs agree to at least two decimal places for the a's, b's, and c's. Figure 10 shows the residuals of the item parameter estimates. The estimates of item discrimination are clearly improved when compared to Figure 5. There is less scatter, although there is more overestimation. The estimates of item difficulties are also clearly improved, although the correlation of errors of estimates that induces the bias in the estimated difficulties is still present. One of the reasons for the introduction of the four-step procedure as the standard for LOGIST

BEST COPY AVAILABLE

runs was to prevent the overinteraction between item and person parameter estimates that can cause some estimated a's to increase without limit (Wingersky, private communication, 1988). It seems plausible that the results obtained here are due to the kind of overinteraction that the four-step procedure was designed to prevent.

Insert Figure 10 about here

Test C1 and BILOG

BILOG approaches the problem of obtaining estimated item and person parameters in two phases. First, item parameter estimates are obtained using MML procedures, assuming a distribution of ability and allowing the imposition of formal prior distributions on the item parameter estimates. Second, after the item parameters have been obtained, ability estimates may be obtained by a number of different estimation procedures, assuming the item parameter estimates are the true values. Given the behavior of LOGIST's implementation of the JML procedures with these extreme tests, it is of interest to investigate the properties of BILOG for at least one of these extreme tests.

The first BILOG run to obtain item parameter estimates used the response data for the calibration sample for test C1, which has a positive correlation of .8 between true a and b. The distribution of abilities was specified as rectangular, and fixed for the run. The starting values for all difficulties were set to 0.0 and the starting values for all discriminations were set to 1.0; default prior distributions on the item parameters were used. This run failed to converge, in fact, it diverged to unreasonable values for some of the item parameter estimates. During the estimation cycles, some item discriminations became strongly negative, for example, equal to -2.

Mislevy (personal communication, 1988) suggested that the process could be improved by specifying that initial values for item discriminations and difficulties be computed hueristically from the conventional proportions correct and r-biserials for each item. This method of obtaining initial values is the BILOG default option, done automatically unless specifically bypassed, as in the first BILOG run. A second BILOG run was done following this suggestion, still retaining the default prior distributions on the item parameter estimates and the fixed rectangular distribution of ability. This run came to a successful conclusion. It is disturbing to find that BILOG also is

apparently sensitive to the starting values used for the parameters.

In the second phase of BILOG, maximum likelihood estimates of ability were computed for the calibration sample assuming that the item parameter estimates are the true values. The calibration sample serves the same function here for BILOG that the cross-validation sample did for LOGIST in that the estimated abilities for the calibration sample played no direct role in the estimation of item parameters and are estimated from only the final item parameter estimates.

The residuals of the item parameter estimates are shown in Figure 11, and the residual median ability estimates in Figure 12. These may be compared to the typical LOGIST results for test C1 shown in Figure 4 (for item parameter estimates) and Figure 3 (for ability estimates). BILOG estimated discriminations have less scatter than typical LOGIST estimated discriminations of Figure 4. The estimated difficulties are clearly less biased than those of LOGIST in Figure 4. However, the correlations between errors of estimates of a and b still cause some bias in estimated difficulties, which cause the estimated abilities to exhibit bias also. The degree of this bias appears to be approximately the

same as when LOGIST was run to complete convergence on test C2 (Figures 9 and 10).

Three More Realistic Tests

The four tests of the previous section were designed to be extreme in order to illustrate certain phenomena. It is reasonable now to consider whether these same phenomena occur when the set of items to be calibrated is more similar to tests that are actually constructed and administered in practice. To investigate this issue, the results of an actual LOGIST calibration of 130 5-choice SAT Verbal items were used as a basis for constructing three new tests, summarized in Table 2, as follows:

1) Test T1 was designed to be a typical test; the true item parameters were defined to be the same as those obtained from the calibration of the SAT Verbal form. The overall correlation between the true item discriminations and true difficulties was .27. For the 32 items with true difficulties less than -1, the correlation between true a and b was .50; for the 51 items with true difficulties between minus and plus 1.0, the correlation between true a and b was .11; and for the 47 items with true difficulties greater than 1., the correlation between true a and true b was -.15.

2) Test T2 was designed to be unusually discriminating. The true item difficulties and guessing parameters were the same as for the typical test T1. The true item discriminations were obtained by multiplying each discrimination parameter in the typical test by 2.

3) Test T3 was designed to be a poorly discriminating test. The true difficulties and guessing parameters were the same as the typical test T1. The true item discriminations were obtained by dividing each discrimination parameter in the typical test by 2.

Insert Table 2 about here

The calibration sample and the cross-validation sample from the previous simulations were used in these simulations also. As before, item response data were generated for each of these new tests for each simulee in each sample. The residual median plots for the cross-validation sample when ability estimates are computed from the generated response data and the true item parameters for each test are shown in Figure 13. The residual medians have roughly the same shape, although there is more variation in these residuals for the poorly discriminating test T3, than for the other two tests. The width of the confidence

bands is narrower for the discriminating test T2 than for the typical test T1, and it is broader for the poorly discriminating test T3 than for the typical test T1. This is as expected, since the standard error when estimating ability from true item parameters is a function of item discriminations (Lord, 1983a). As before, when compared to the analagous plots for the extreme tests, shown in Figure 2, the predicted statistical bias is insignificant.

Insert Figure 13 about here

Each test was calibrated using the standard LOGIST procedure, and the results transformed to the metric of the true item parameters. The transformed item parameter estimates, along with the generated data for the cross-validation sample were then used to compute estimated abilities. The results, in terms of the residual median plots, are shown in Figure 14 for each test. It is again well to remember that the calibration sample of simulees was the same for these three tests as well as the four extreme test studied earlier, as was the cross-validation sample.

Insert Figure 14 about here

The residual medians for all three tests show biases similar to, but smaller than, those of extreme test C1, which had a correlation of +.8 between true discrimination and true difficulty. Here the correlation between true discrimination and difficulty is .27. The residual medians are most biased for the most discriminating test, T2, and least biased for the least discriminating test T3. The typical test lies inbetween these two. The width of the confidence bands is a function of the estimated item discriminations and is widest for the poorly discriminating test T3, and narrowest for the highly discriminating test T2. It is sufficiently wide for test T3 that one cannot conclude that these residual medians are significantly biased at most levels of true ability. Apparently, if a test is of sufficiently low discrimination, there is enough error in the ability estimates for the cross-validation sample so that bias in the ability estimates cannot be easily detected.

Figures 15, 16, and 17 show the residuals for the estimated item parameters for the three tests. The residual plots for the item difficulties are on a different scale than those for the

extreme tests discussed in the previous section in order to accommodate the wider range found in the more typical tests here. Note also that the scale on which the item discrimination residuals are plotted differs from that of the extreme tests, and also across the three tests of the present section. For the typical and discriminating tests, the overall patterns of bias for the item difficulties is similar to that of the estimated abilities, although the pattern is somewhat clearer for the more discriminating test. For these two tests, the correlations between errors of estimates of a and b for hard and easy items are visible. For the poorly discriminating test T3, Figure 17, the scatter in estimated difficulties is much wider than we have seen. This Figure shows that when item discriminations are very low, if they are underestimated, item difficulty is also, regardless of the true difficulty. Conversely, when low discriminations are overestimated, item difficulties are also, regardless of the value of the true difficulty.

Insert Figures 15, 16, and 17 about here

The scatter shown for the residuals for item discriminations differs for the three tests. The a parameters for the poorly

discriminating test T3 are the most precisely estimated; the a parameters for the typical test T1 are estimated with moderate precision, and the a parameters for the most discriminating test T2 are estimated with the least presicion. This is in accord with the Wingersky and Lord (1984) result that, when item and ability parameters are jointly estimated, the standard error of a increases with a. It is clear that the guessing parameter is not very well estimated for items with low discrimination. The majority of items have guessing parameters estimated to be the common value by LOGIST for the poorly discriminating test T3. This does not happen for either the typical test or the discriminating test.

Tests T1 and T2 and LOGIST

The typical test T1 and the more discriminating test T2 show the same type of biases as the more extreme test C1 studied earlier. They present an opportunity to confirm again the hypothesis concerning the sensitivity of the results of the typical LOGIST calibration to the starting value of the discrimination parameters. To test the hypothesis in this context, the calibrations of these two tests were repeated, using the true item discriminations as the starting values for the estimated discriminations.

The results of these calibrations in terms of the residual medians for the calibration (not cross-validation) sample are shown in Figure 18. As expected, the residual medians show very little, if any, of the bias seen in Figure 14. Likewise the residuals for the item parameter estimates shown in Figures 19 and 20 show the correlations between errors of estimates of difficulty and discrimination to be much reduced, although still visible for the most extreme items.

Insert Figures 18, 19, and 20 about here

An Unusual Set of Items in a Realistic Setting

The Department of Defense has engaged in a number of coordinated efforts over the past decade aimed at exploring the feasibility of replacing the Armed Services Vocational Aptitude Battery (ASVAB) with a computerized adaptive battery. The ASVAB is administered to all candidates for military service for the purposes of both selection and placement. As part of the exploration of adaptive testing, Prestwood, Vale, Massey and Welsh (1985) developed and calibrated over 2000 experimental items to be considered as candidates for an ASVAB adaptive testing item pool. In a recent effort to develop methods of on-line calibration,

funded by a number of agencies of the Department of Defense, the author, along with three other researchers, made an extensive exploration of the subset of these items developed for the Word Knowledge (WK) subtest (Stocking, 1988a; Holland, Bock, Davis, Levine, Samejima, & Stocking, 1988). The focus of the study described here is the simulation of the initial calibration of these experimental WK items as candidates for an adaptive testing item pool.

The current WK subtest consists of a single item type, synonyms, and is designed to measure the understanding of words typically used in social studies and everyday life, human relationships, science and nature, and arts and humanities (Prestwood, et al., 1985). Prestwood et al. wrote and calibrated 258 similar items that spanned a wider range of difficulty, as required by adaptive testing, than those found in operational use. For the purpose of obtaining item parameter estimates they acquired a sample of N = 8171 candidates for military service from Military Entrance Processing Stations who had also taken the conventional ASVAB test battery. This set of data, that is, the responses of 8171 individuals to the 'experimental' Word Knowledge items as well as the conventional ASVAB, forms the basis of the current simulation.

The set of items is unusual in that, while each of the current operational forms is presumably designed to measure the typical or average examinee most efficiently, the experimental items are designed to span a much broader range. Two aspects of the obtained set of examinees deserve mention. First, it is unlikely that motivation was the same on the experimental items as on the operational items, since examinees volunteered responses to the experimental items while being required to respond to the operational forms for military service entrance. Second, the true distribution of ability for this sample is undoubtedly bell-shaped, perhaps providing less than optimal information for estimating parameters of items that span a wide range of difficulty.

For the purpose of collecting response data, the 258 items were divided into three roughly parallel 'tests' of 86 items each. Each candidate took one of these experimental tests, in addition to one of six different 35-item WK subtests in current operational use. These operational subtests serve as links in the calibration design in the sense that individuals taking the same operational subtest were assigned to be administered any one of the three experimental tests. The resulting sparse data matrix consists of responses to 468 items: 3 experimental tests of 86 items each

plus 6 operational subtests of 35 items each. Each experimental test was taken by at least 2500 examinees, and each subtest of an operational form was taken by at least 1100 examinees.

Item parameter estimates were obtained by Prestwood et al. for all items in a single calibration using the ASCAL (Vale & Gialluca, 1988) procedure. For the present simulation, these parameter estimates are considered to be the true item parameters. Maximum likelihood estimates of examinee ability were computed, using the true item parameters, and these ability estimates are considered to be true abilities for the simulees in the current study. Summary statistics for these true item and person parameters are given in Table 3. The correlation between true discrimination and true difficulty for the 468 items is .48. For the 218 items with true difficulty less than -1., the correlation between discrimination and difficulty is .43; for the 202 items with true difficulty between -1. and +1., the correlation is .11, and for the 48 items with true difficulty greater than +1., the correlation is -.27. This pattern of correlation coefficients is similar to the .50, .11, -.15 pattern seen for the easy, moderate, and harder items in the three tests of the previous study.

Insert Table 3 about here

Simulated response data were produced, in the same sparse matrix design as the real data, using the true item parameters, the true simulee abilities, and the 3PL item response function model. Maximum likelihood simulee ability estimates were then obtained from the generated data, using the true item parameters. The resulting residual median plot is shown in the top panel of Figure 21. Note that horizontal axes in all panels of Figure 21 have a wider scale than all other such plots in this paper, to accomodate the wider range of ability found in this sample of simulees. The confidence bands are wider at the extremes than in other plots of this nature because of the smaller number of cases in the tails of the ability distribution. All simulees with true ability less than -3 or greater than +3 are grouped into the lowest and highest points plotted. This avoids the extreme fluctuations that might otherwise occur because of very small numbers of cases in the most extreme tails. As before, the statistical bias of these ability estimates appears negligible.

Insert Figure 21 about here

The entire set of 468 items was calibrated in a typical LOGIST run, and the results transformed to the metric of the true item parameters. The results, in terms of the residual median

plot for the abilities of the calibration sample, are shown in the middle panel of Figure 21. Figure 22 shows the residuals for the estimated item parameters, on scales different from other such residual plots in this paper to accomodate these data. As with the other items sets that had positive correlations between true a and true b, namely C1, T1, T2, and T3, extreme abilities are overestimated. Extreme difficulties are also overestimated, and the correlation between estimation errors for item discrimination and difficulty is strong, particularly for easy items. This is due, in part, to the fact that the particular sample of simulees contained few individuals with extreme abilities and therefore little information for estimating the parameters of extreme items. The mean true item discrimination for this item set is 1.26, which is fairly high. As with the artificial discriminating test T2 of the previous study, the estimates of item discrimination show a fair amount of scatter and are, on average, overestimates.

Insert Figure 22 about here

To confirm some of the results obtained previously, this calibration was repeated, reading in the true discriminations parameters as starting values. The residual median plot for the abilities of the calibration sample is shown in the bottom panel

of Figure 21, and the residuals for the item parameter estimates are shown in Figure 23. The bias for the estimated abilities is much reduced. The bias for the item difficulties is also reduced, but the correlation between errors of estimation for item discriminations and difficulties is still visible. Item discriminations are more overestimated here than when a common middle value is used as the starting value for all item discriminations.

Insert Figure 23 about here

Conclusions and Recommendations

The motivation underlying this research was to explore and understand some apparently anomalous results in various LOGIST-based applications of IRT that have been obtained from time to time over the past several years (see, for example, Lord, 1975b; Stocking, Cook, & Eignor, 1988; Stocking 1988a). At the time, some of these anomalous results were attributed, at least in part, to the fact that, in reality, the 3PL item response function never fits real data exactly (Eignor & Stocking, 1986). But other anomalous results were obtained in simulation studies, such as this one, where data are generated to fit the 3PL model. The applications themselves are unimportant for the present

discussion, but had one element in common, that is, they all depended heavily upon the use, in contexts perhaps unrelated to the actual calibration of items, of item parameter estimates as if they were true parameters. While this unavoidably introduces errors in the subsequent uses of the items, it was also found to introduce large (and sometimes unacceptable) biases.

On the basis of the research presented here, it seems clear that theoretical (statistical) bias can be neglected; it also seems clear that empirical bias cannot be neglected. The typical LOGIST implementation of the JML procedure gives rise to some of the anomalous results previously obtained. The structure of the four-step procedure imposed on the alternating stages of ability and item parameter estimation can sometimes prevent sufficient movement away from starting values, consequently preventing the production of 'good' final estimates. And the nature of the biases in the final estimates obtained is a function of the true properties of the item sets calibrated, as seen in Figures 3 through 7.

This can be clearly seen when true values of item discrimination are used as starting values compared with the typical middle value for all item discriminations, as in Figures 8 and 9; 18, 19 and 20; and 21, 22 and 23. The four-step procedure does not move very far away from the starting values, therefore

using true discriminations as starting values produces much better results.

The best a JML-based estimation procedure can hope to do is exemplified by running LOGIST to complete convergence, as in Figures 9 and 10. Although extra monetary costs are incurred in this setting, starting values for the parameters are irrelevant. The best an MML-based estimation procedure can hope to do is exemplified by running BILOG to convergence with good starting values, as in Figures 11 and 12. Poor starting values can lead to divergence of the procedure. However, and this bears emphasis, even in these most optimum settings, such as rectangular calibration samples and completely appropriate items, the naturally occurring correlations among errors of estimation for the item parameters, one of which is shown in Figure 1, do not disappear; they remain, although their deleterious consequences are reduced.

Stocking (1988b) shows why such correlations are inevitable. Examinees with ability equal to item difficulty are optimum for estimating only the item difficulty, and provide little information for estimating either item discrimination or the guessing parameter. Only examinees whose true ability lies above and below the item difficulty provide information for estimating item discrimination; and only examinees whose true ability lies

far below the item difficulty provide information for estimating the guessing parameter. It seems likely that there will always be, in any item set/calibration sample combination, some extreme items for which the calibration sample cannot provide adequate data. And it is these items for which estimated parameters will be in most error and will exhibit most strongly the correlations among the errors. Whether these correlations remain important in the subsequent use of item parameter estimates, assuming they were true, is an issue that must be evaluated carefully in the context of the particular application.

It seems likely that the future holds some promise of improved estimation methods that may mitigate some of the problems described in this paper. Methods that do not rely solely on point estimtes of parameters, but rather work from their posterior distributions, may potentially provide better results. These methods (e.g., Tsutakawa and Soltys, 1988), formally incorporate sources of uncertainty--including the error correlations that play a central role in the present paper--contained in the estimates through Bayesian methods.

LOGIST, as a computer program of wide and long-standing use in many different applications of IRT, needs improvement. Most applications cannot afford to run the program to complete convergence. It may be possible to improve results of the

BEST COPY AVAILABLE

four-step structure by obtaining better starting values for the parameters. Alternatively, controlling the behavior of estimates of discrimination and guessing parameters through the imposition of prior distributions on them may be cost effective and provide reasonable results.

BILOG, being a more recent computer program available for general use, has not been subjected to the same wide variety of applications as LOGIST. As such, it does not contain the necessary restrictions to prevent the numerical procedures from diverging from reasonable, although perhaps less than optimal starting values. It seems clear that such additional restrictions are necessary.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. (pp. 397-479). Reading MA: Addison-Wesley.

Cook, L. L., Petersen, N. J., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 136-156.

Davey, T., & Levine, M. (1988). Improved sampling variances of parameter estimates in item response theory models. Paper presented at the annual ONR Contractor's Conference, Iowa City, IA.

David, H. A. (1981). Order statistics (2nd ed.). New York: Wiley.

Eignor, D. R., & Stocking, M. L. (1986). An investigation of possible causes for the inadequacy of IRT pre-equating (Research Report 86-14). Princeton, NJ: Educational Testing Service.

Holland, P. W., Bock, R. D., Davis, C. E., Levine, M. V., Samejima, F., & Stocking, M. L. (1988, in preparation). Final Report on on-line calibration.

Lewis, C., & Sheehan, K. M. (1988, in press). Computerized mastery testing. Machine Mediated Learing.

Lord, F. M. (1975a). The 'ability' scale in item characteristic curve theory. Psychometrika, 40, 205-217.

Lord, F. M. (1975b). Evaluation with artificial data of a procedure for estimatign ability and item characteristic curve parameters. (Research Report 75-33). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983a). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245.

Lord, F. M. (1983b). Statistical bias in maximum likelihood estimators of item parameters. Psychometrika, 48, 425-435.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. Journal of Educational Measurement, 2, 157-162.

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference.

Minneapolis, MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item analysis and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Stocking, M. L. (1987). A consumer's guide to LOGIST and BILOG (Research Report 87-43). Princeton, NJ: Educational Testing Service.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). Armed Services Vocational Aptitude Battery: Development of an adaptive item pool (AFHRL-TR-85-19). Brooks Air Force Base, TX: Manpower and Personnel Division, Human Resources Laboratory.

Stocking, M. L. (1987). Two simulated feasibility studies in computerised adaptive testing. Applied Psychology: An International Review, 36, 263-277.

Stocking, M. L. (1988a). Scale drift in on-line calibration (Research Report 88-28-ONR). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1988b). Optimal examinees for item parameter estimation (Research Report 88-00). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Eignor, D. R. (1986). The impact of different ability distributions on IRT pre-equating (Research Report 86-49). Princeton, NJ: Educational Testing Service.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). Factors affecting the sample invariant properties of linear and curvilinear observed and true score equating procedures (Research Report 88-41). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximation for Bayesian ability estimation. Journal of Educational Statistics, 13, 117-130.

van der Linden, W. J., & Boekkooi-Timminga, E. (1987). A maximin model for test design with practical constraints (Research Report 87-10). Enschede, The Netherlands: University of Twente. To appear in Psychometrika.

Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. Applied Psychological Measurement, 12, 53-68.

Weiss, D. J. (Ed.) (1975). Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN:

Psychometric Methods Program, Department of Psychology, University of Minnesota.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 453-461.

Table 1

Statistics summarizing the true item parameters of the four extreme tests, C1, C2, C3, and C4. All true c's are equal to .15; all tests have 100 items.

Test C1, r = +.8

	mean	S.D.	min	max	Percentiles 10	25	50	75	90
a	1.45	.39	.51	2.30	.91	1.18	1.47	1.72	1.93
b	-.06	1.39	-2.49	2.39	-1.98	-1.26	-.28	1.14	1.87

Test C2, r = -.8

	mean	S.D.	min	max	Percentiles 10	25	50	75	90
a	1.54	.40	.52	2.46	.98	1.31	1.52	1.83	2.10
b	-.15	1.41	-2.49	2.46	-1.92	-1.40	-.28	.90	1.92

Test C3, r = -.8; +.8

	mean	S.D.	min	max	Percentiles 10	25	50	75	90
a	1.51	.42	.61	2.37	.92	1.19	1.55	1.81	2.03
b	0.00	1.55	-2.47	2.46	-2.25	-1.22	.01	1.54	2.03

Test C4, r = +.8; -.8

	mean	S.D.	min	max	Percentiles 10	25	50	75	90
a	1.47	.40	.61	2.46	.95	1.18	1.49	1.78	1.93
b	-.10	1.48	-2.45	2.46	-2.07	-1.48	.02	1.07	2.03

Table 2

Statistics summarizing the true item parameters of the three more realistic tests, T1, T2, and T3. All tests have 130 items.

					Percentiles				
	mean	S.D.	min	max	10	25	50	75	90
T1,a	.85	.30	.22	1.67	.43	.63	.87	1.04	1.17
T2,a	1.69	.59	.44	3.35	.87	1.26	1.74	2.08	2.34
T3,a	.42	.15	.11	.84	.21	.32	.43	.52	.58
b	.17	1.43	-3.66	2.59	-1.82	-1.00	.60	1.37	1.80
c	.17	.08	.00	.50	.10	.12	.15	.22	.27

Table 3

Statistics summarizing the true item parameters and abilities used in the simulated calibration of ASVAB data.

	ASVAB					Percentiles				
	n	mean	S.D.	min	max	10	25	50	75	90
a	468	1.26	.42	.39	2.43	.75	.95	1.22	1.54	1.82
b	468	-.86	1.46	-3.10	3.01	-2.95	-2.11	-.85	.18	1.02
c	468	.23	.08	.01	.47	.13	.18	.22	.29	.35
θ	8171	-.04	1.06	-7.18	5.05	-1.29	-.67	-.05	.61	1.29

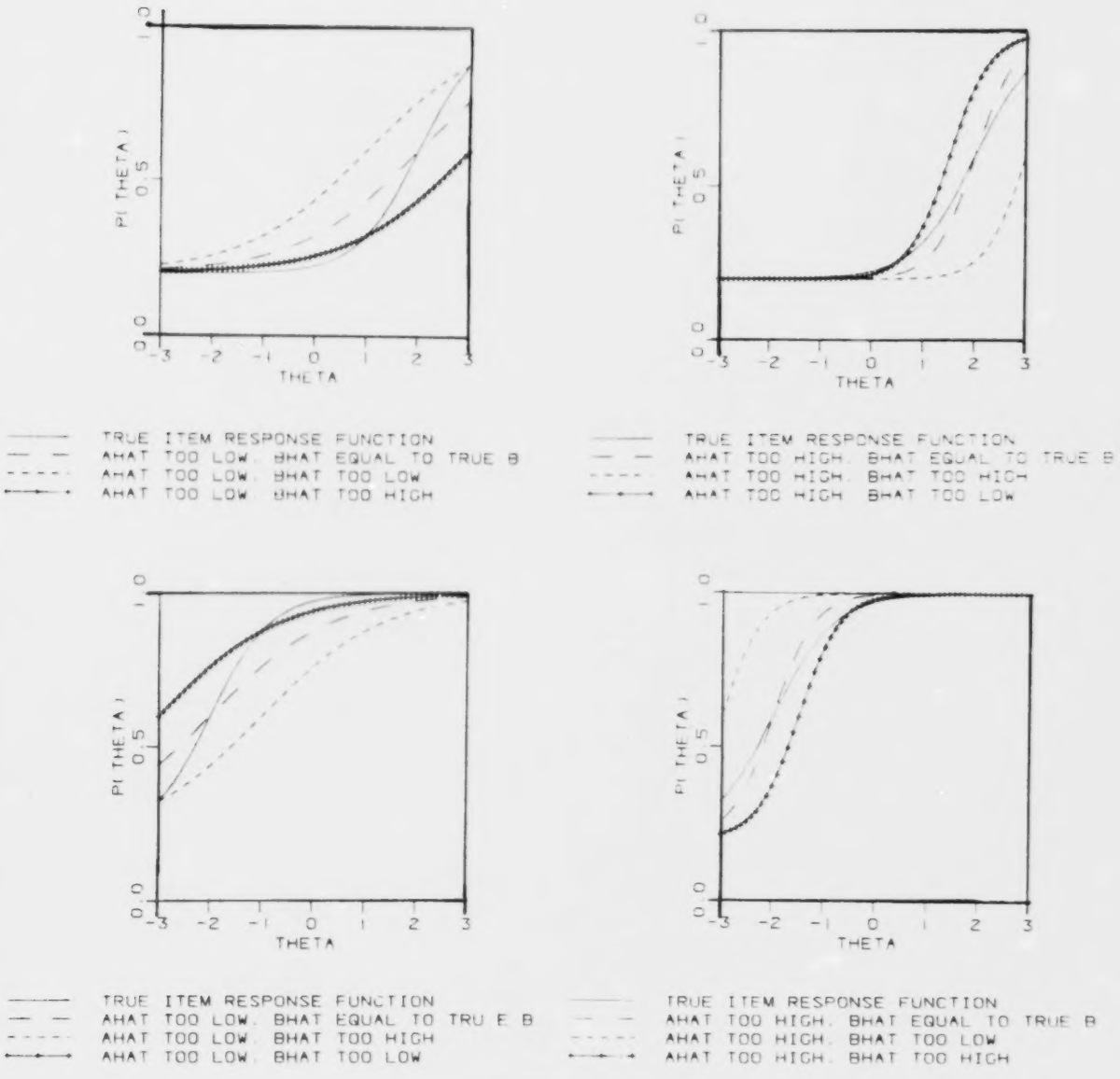

P(THETA)
1.0
0.5
0.0
-3 -2 -1 0 1 2 3
THETA
TRUE ITEM RESPONSE FUNCTION
AHAT TOO LOW, BHAT EQUAL TO TRUE B
AHAT TOO LOW, BHAT TOO LOW
AHAT TOO LOW, BHAT TOO HIGH
AHAT TOO HIGH, BHAT EQUAL TO TRUE B
AHAT TOO HIGH, BHAT TOO HIGH
AHAT TOO HIGH, BHAT TOO LOW

Figure 1: Illustration of the correlation between estimation errors for a and b. The top row shows a hard item; the bottom row shows an easy item. The left column shows the consequences of underestimating the item discrimination; the right column shows the consequences of overestimating the item discrimination.

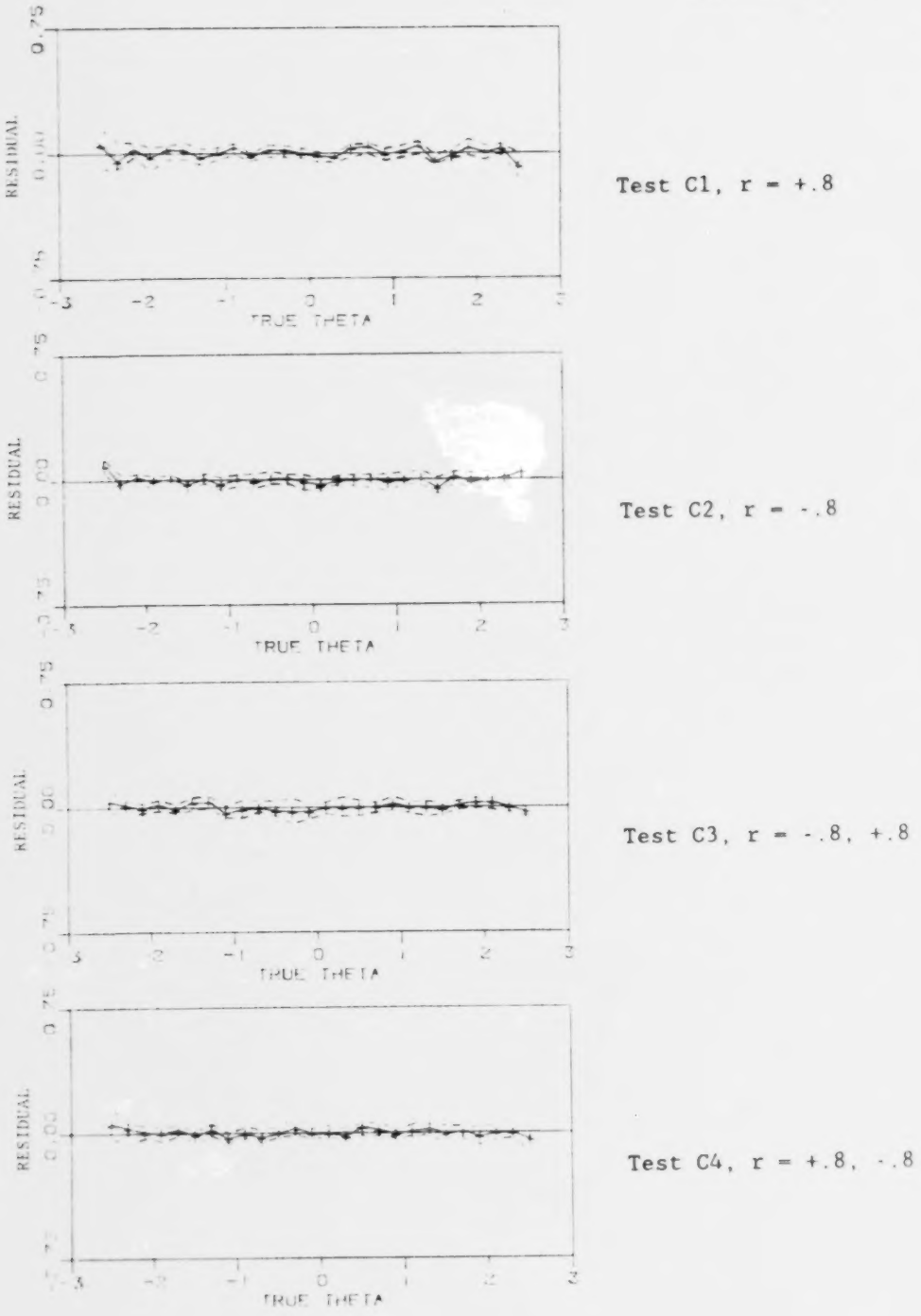
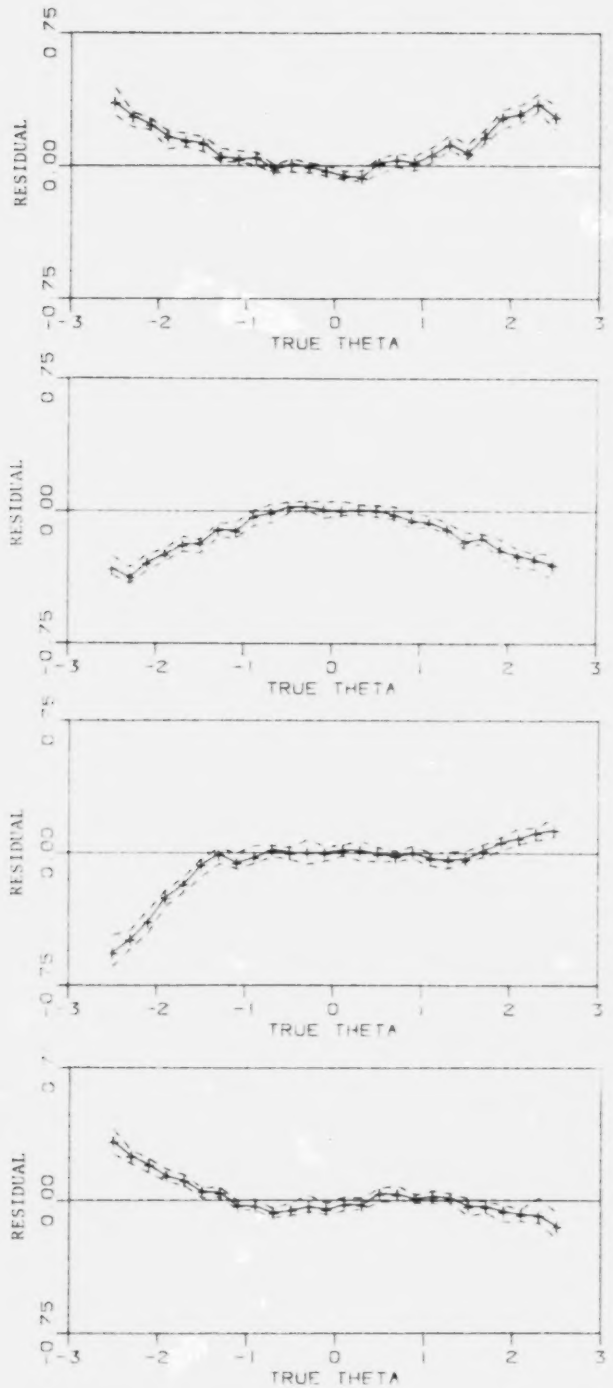


Figure 2: For cross-validation sample abilities estimated from true item parameters, the residual median estimated ability (solid curve) and the 5% two-tailed confidence interval (dashed curves) for the four extreme tests. The tests are ordered with test C1 at the top and test C4 at the bottom.



Test C1, r = +.8

Test C2, r = -.8

Test C3, r = -.8, +.8

Test C4, r = +.8, -.8

Figure 3: For cross-validation sample abilities estimated from estimated item parameters, the residual median estimated ability (solid curve) and the 5% two-tailed confidence interval (dashed curves) for the four extreme tests.

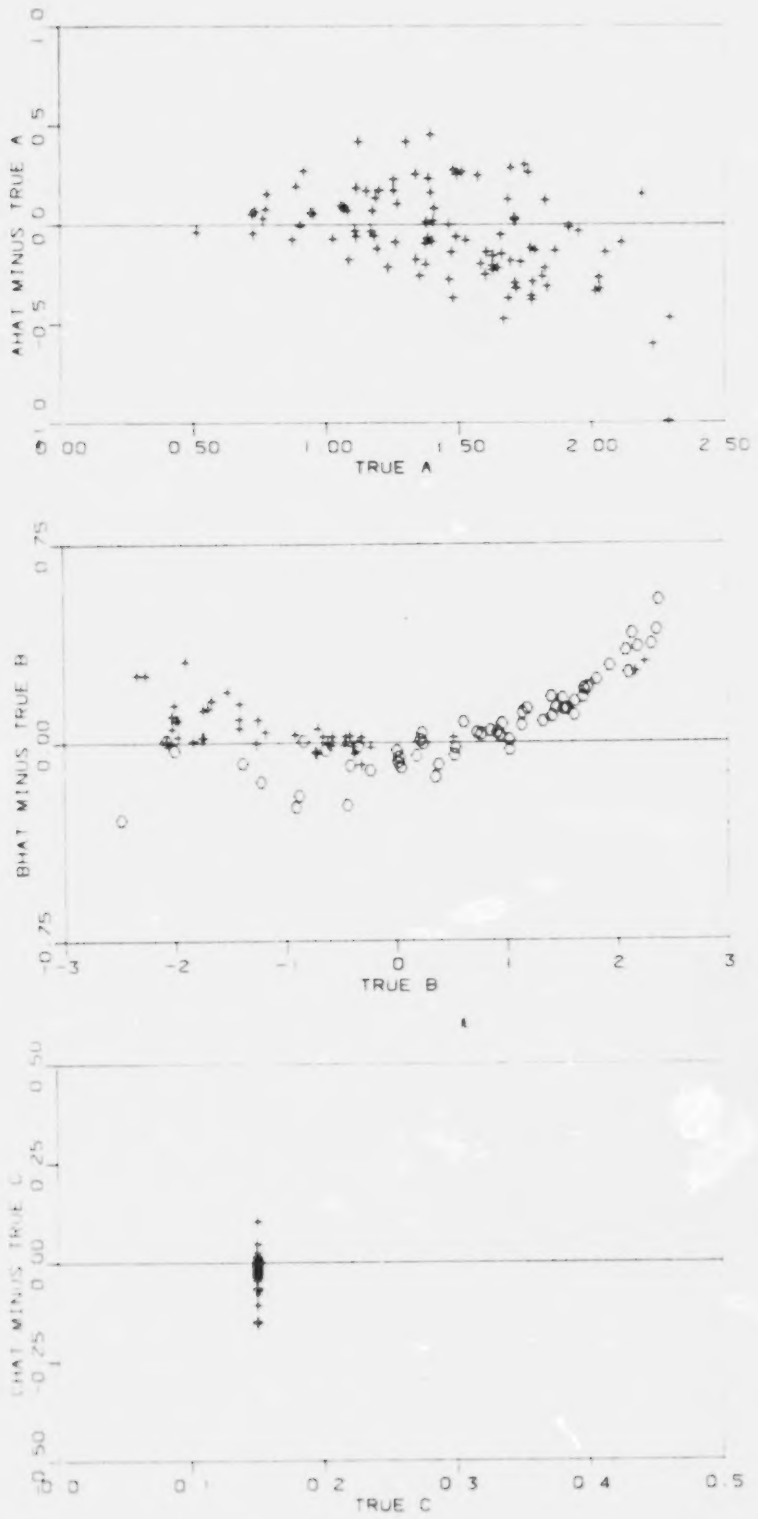


Figure 4: The residuals (estimated minus true) of the estimated item parameters for test C1, r = +.8.

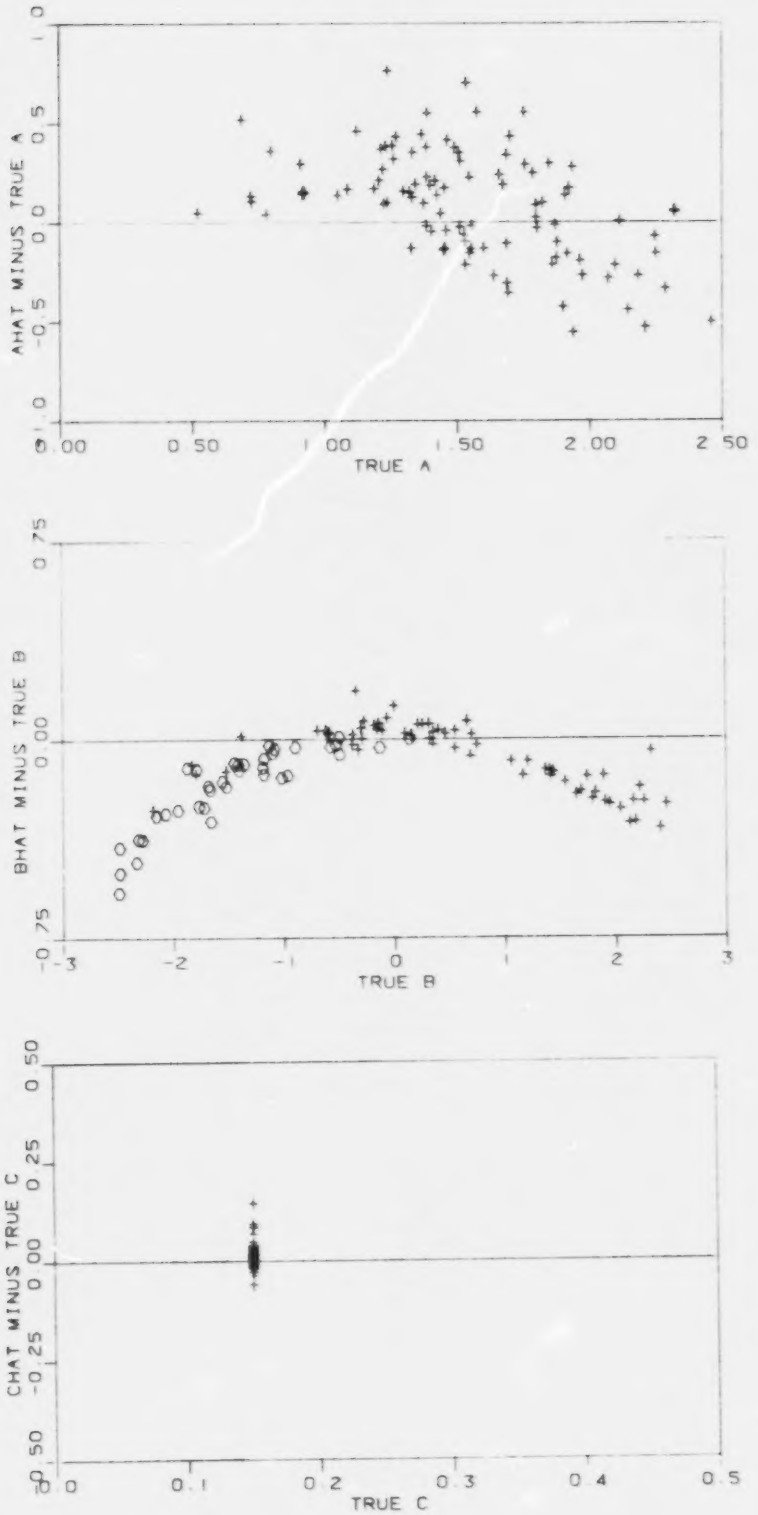


Figure 5: The residuals (estimated minus true) of the estimated item parameters for test C2, r = -.8.

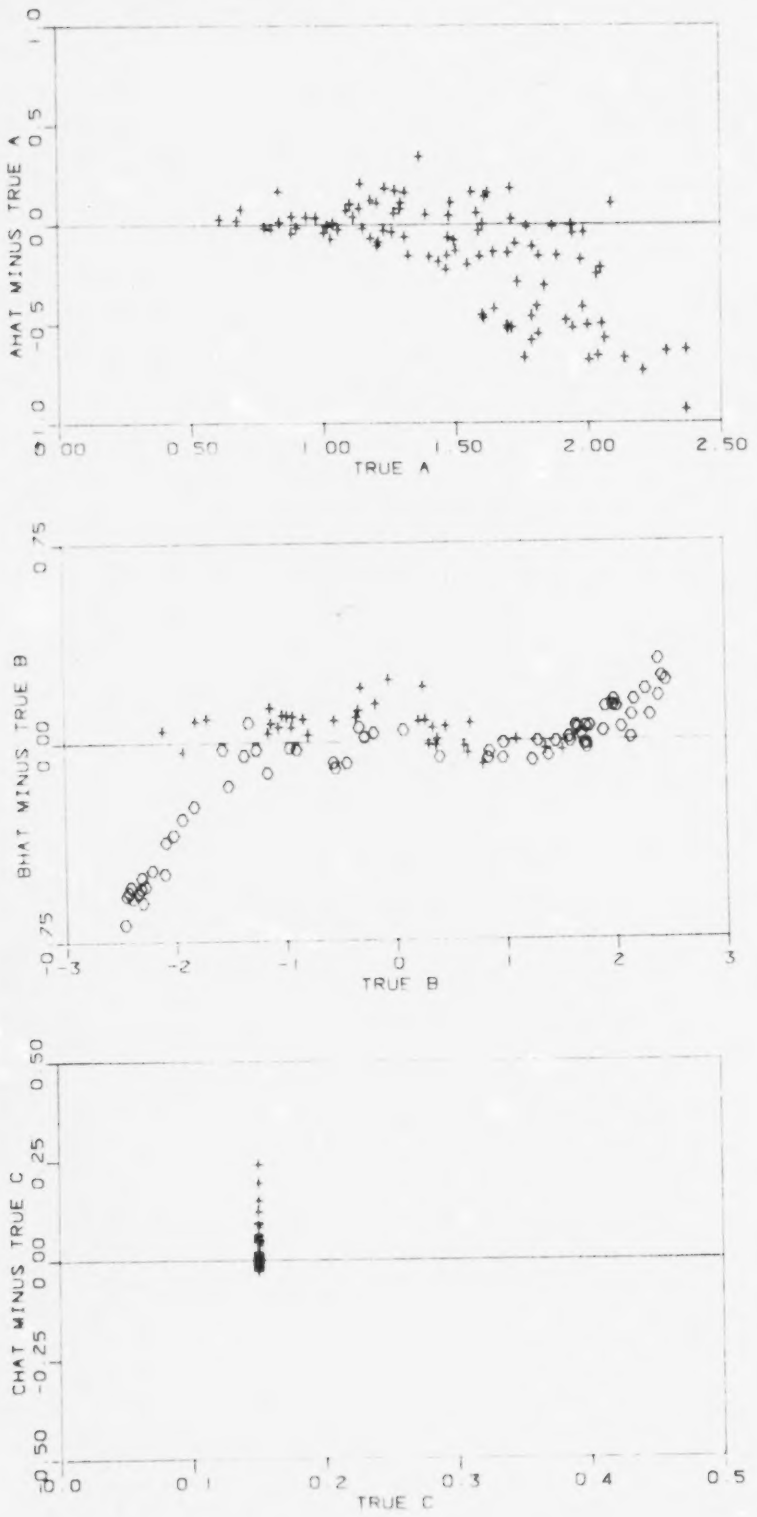


Figure 6: The residuals (estimated minus true) of the estimated item parameters for test C3, r = -.8, +.8.

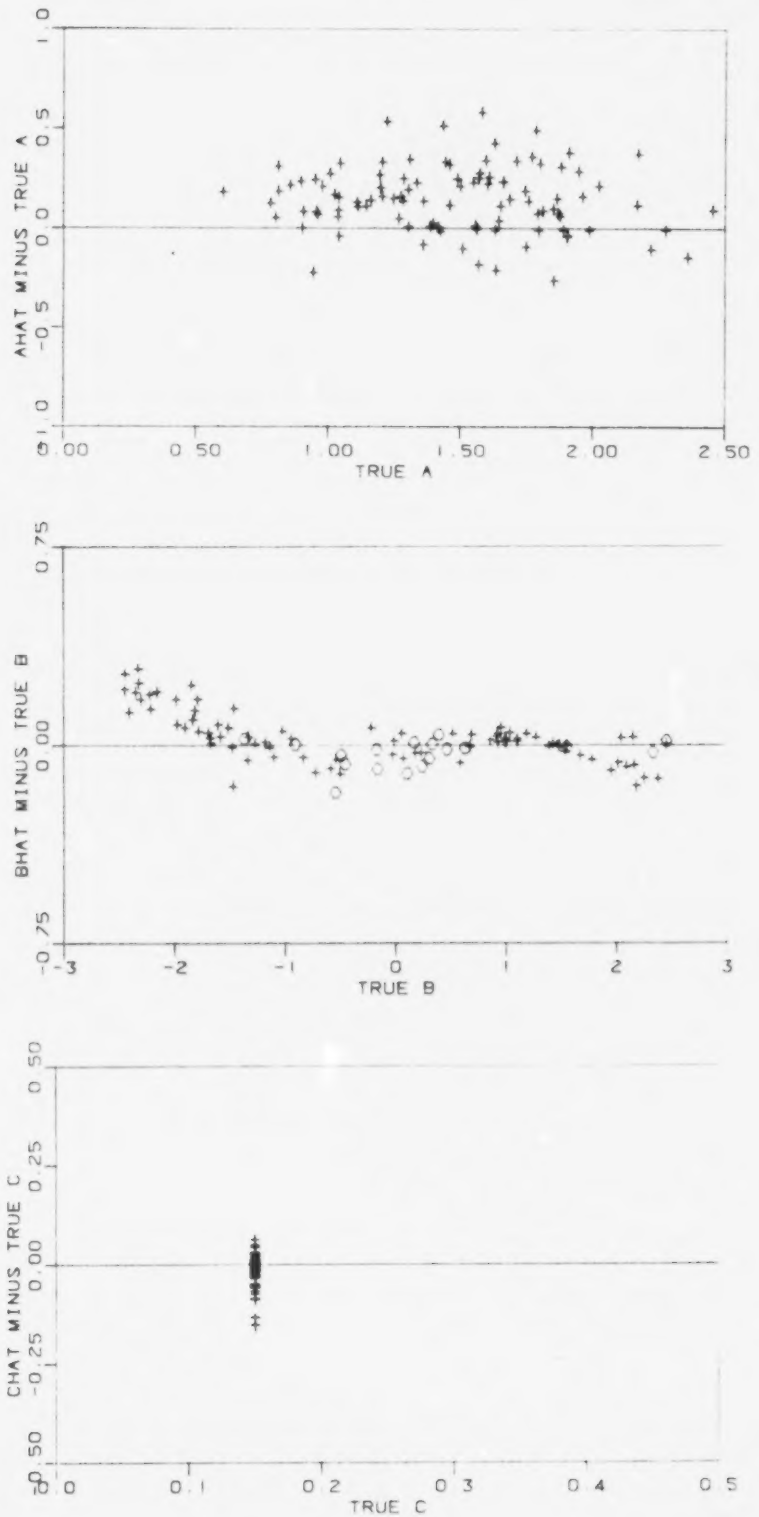


Figure 7: The residuals (estimated minus true) of the estimated item parameters for test C4, r = +.8, -.8.

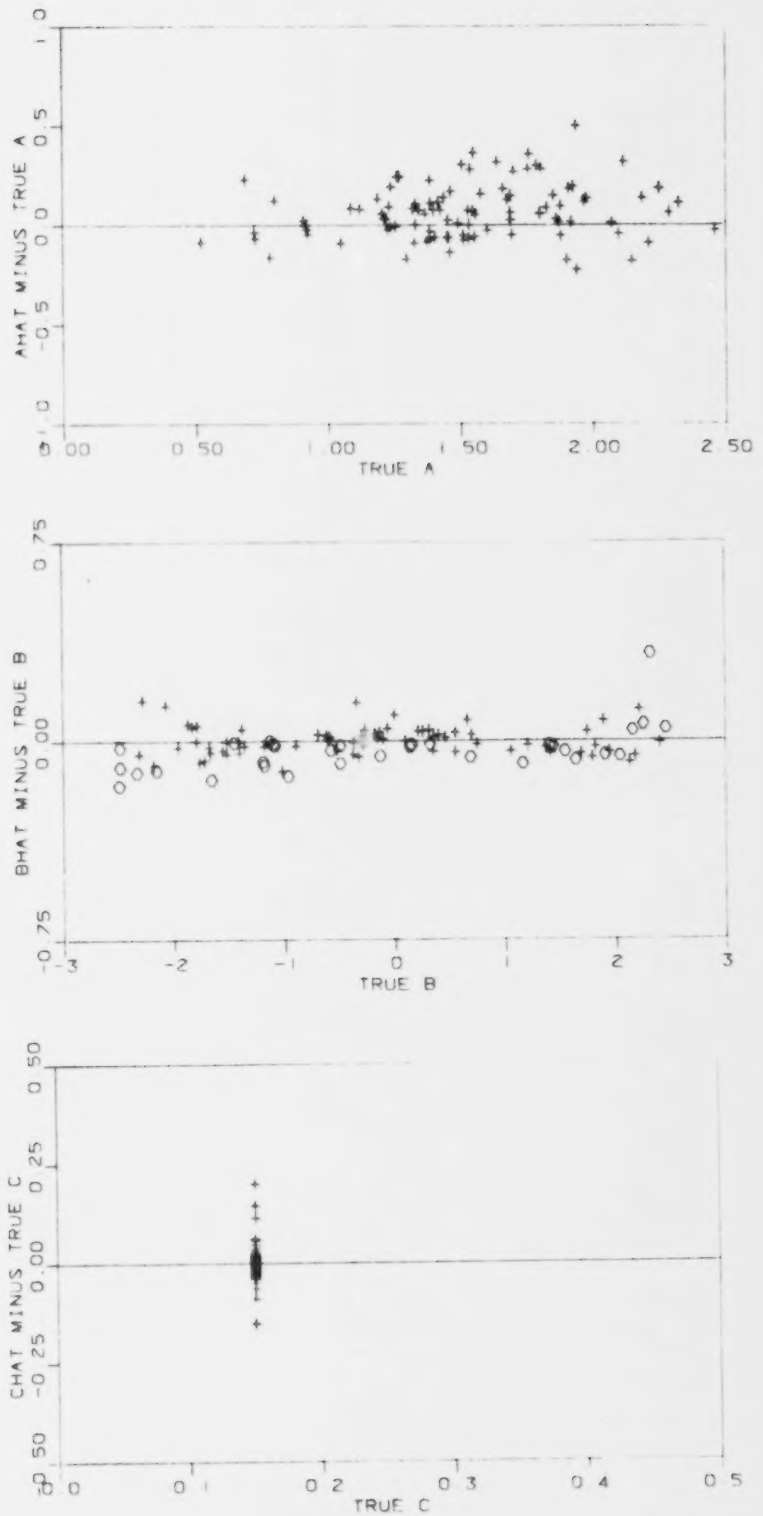


Figure 8: The residuals (estimated minus true) of the estimated item parameters for test C2, r = -.8, when true a's are used as starting values for the item discrimination estimates.

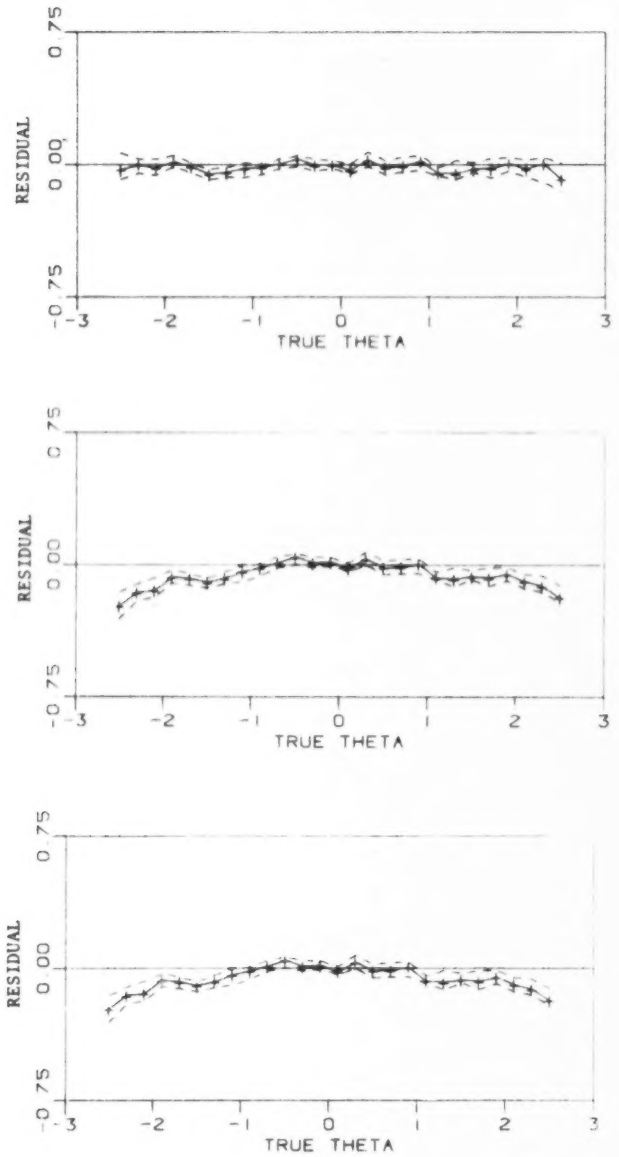


Figure 9: For calibration sample abilities estimated from estimated item parameters, the residual median estimated ability and the confidence interval for test C2, r = -.8, when the LOGIST four-step procedure is used with true a's as starting values (top), when LOGIST is run to complete convergence using a common starting value for a (middle), and when LOGIST is run to complete convergence using the true a's as starting values (bottom).

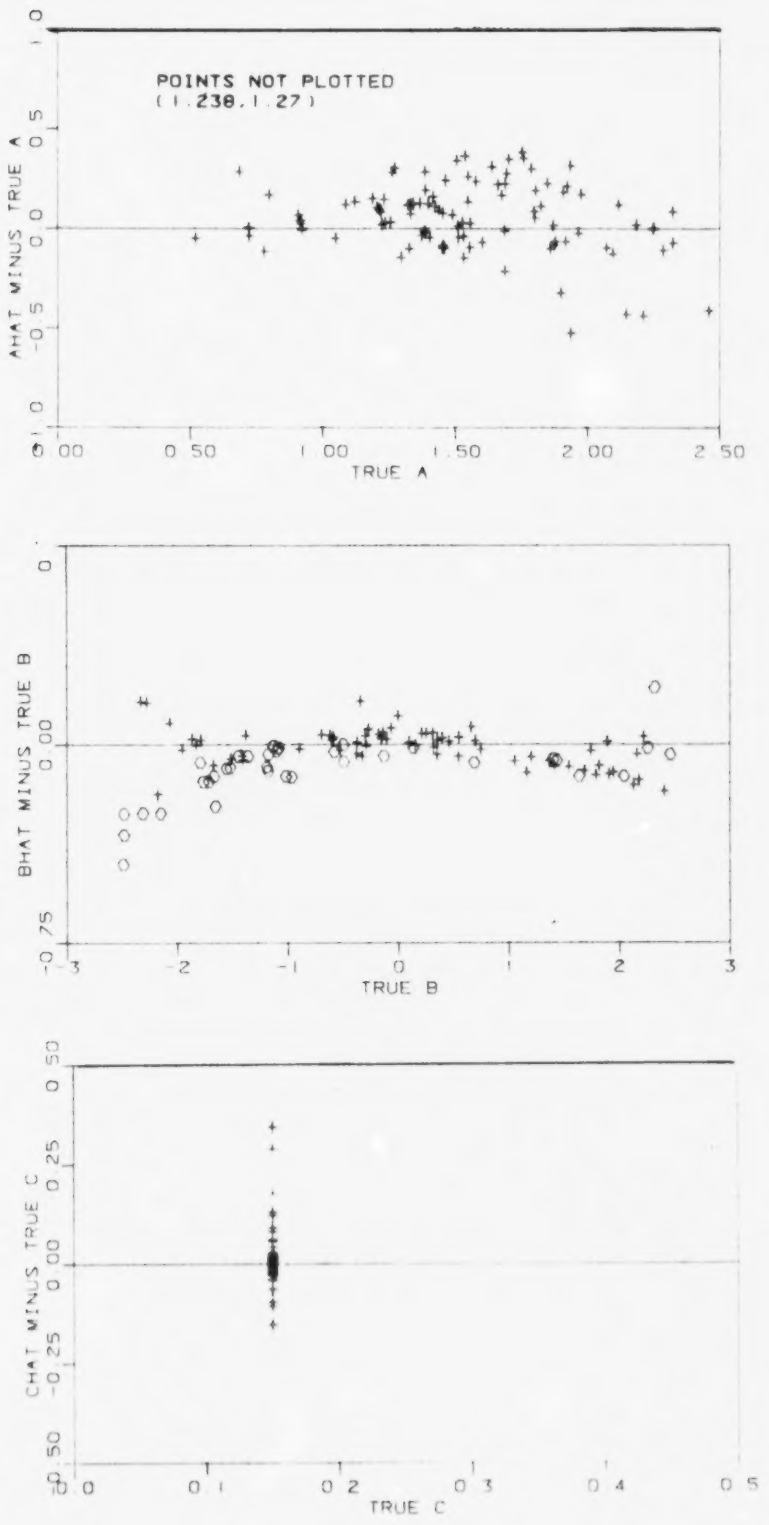


Figure 10: The residuals (estimated minus true) of the estimated item parameters for test C2, r = -.8, when LOGIST is run to complete convergence.

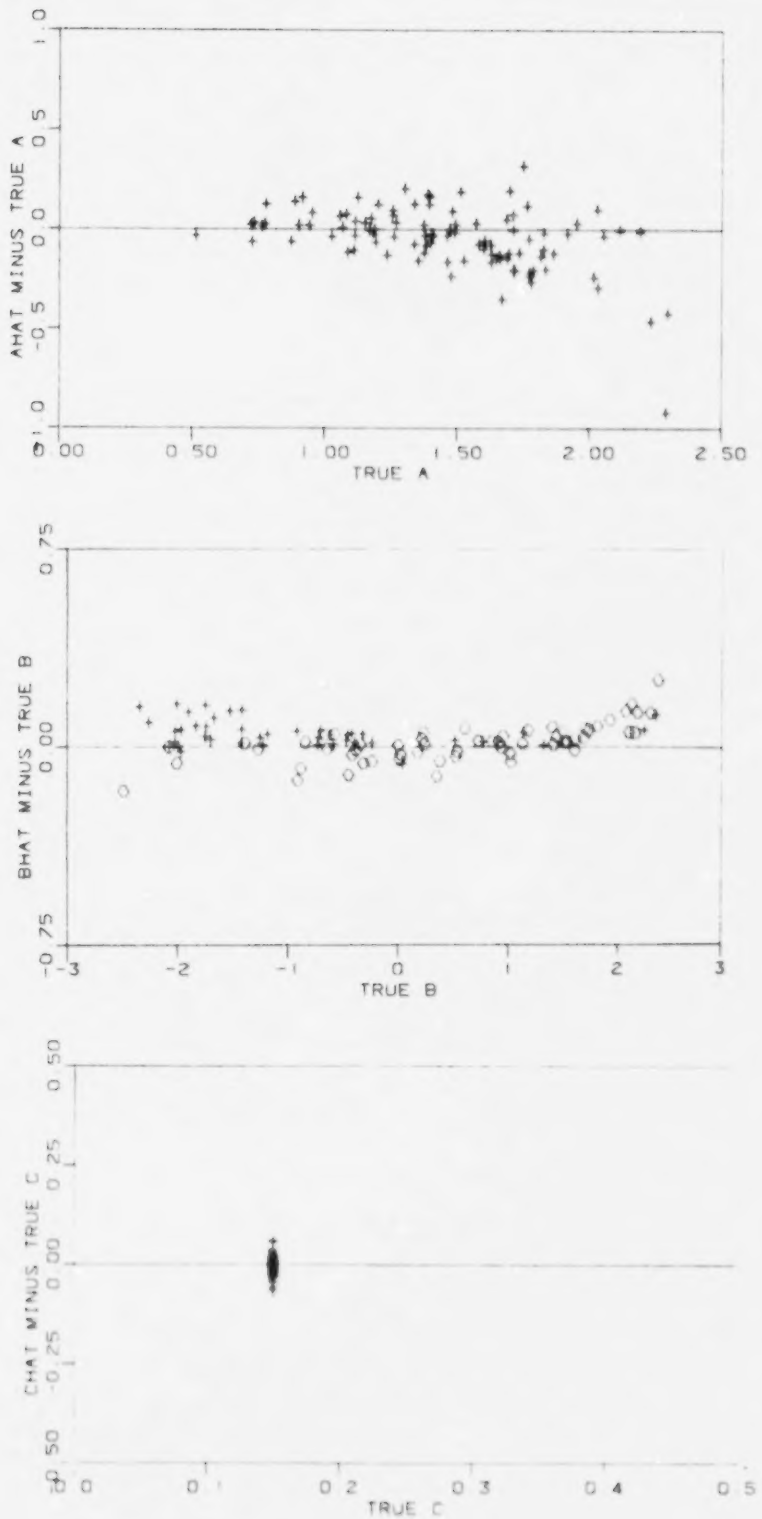


Figure 11: The residuals (estimated minus true) of the estimated item parameters from BILOG for test C1, r = +.8.

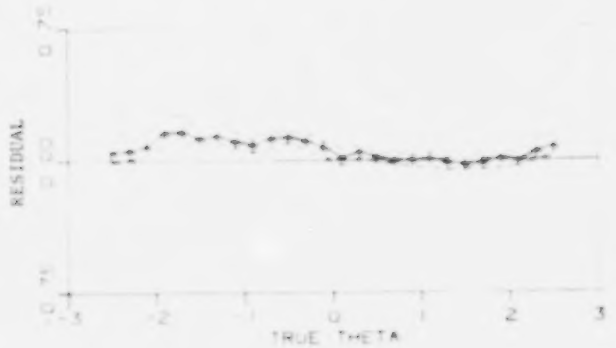
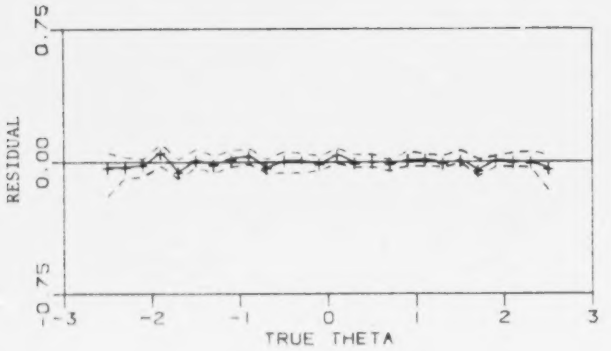
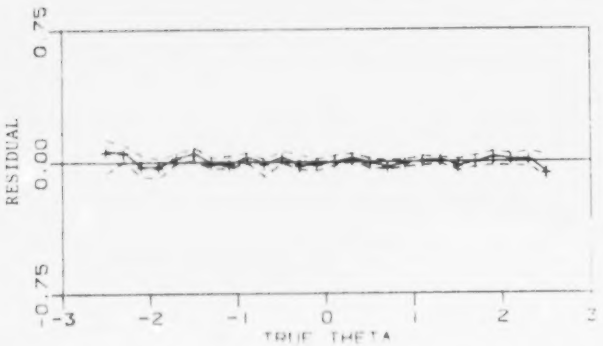


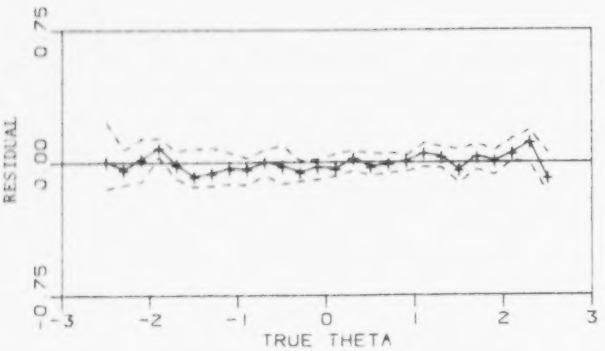
Figure 12: For calibration sample abilities estimated (MLE) from BILOG's estimated item parameters, the residual median estimated ability (solid curve) and the 5% two-tailed confidence interval (dashed curves) for test C1. $r = +.8$.



Typical Test T1

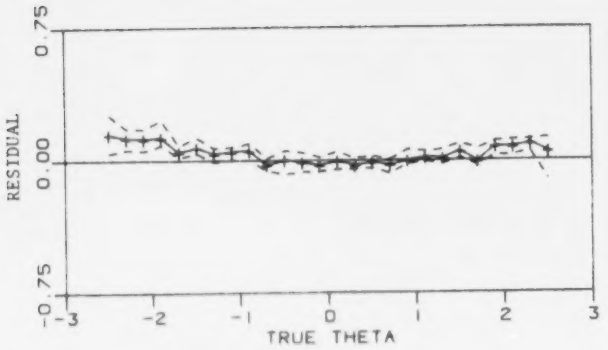


Discriminating Test T2

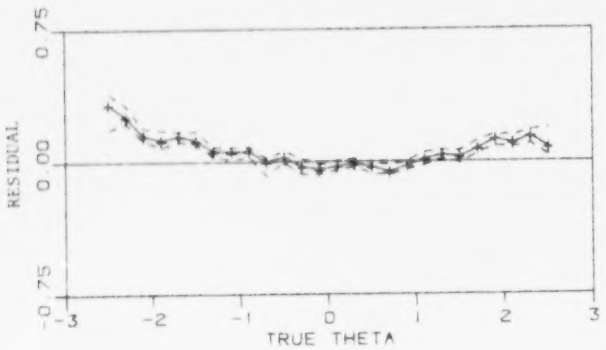


Poorly Discriminating Test T3

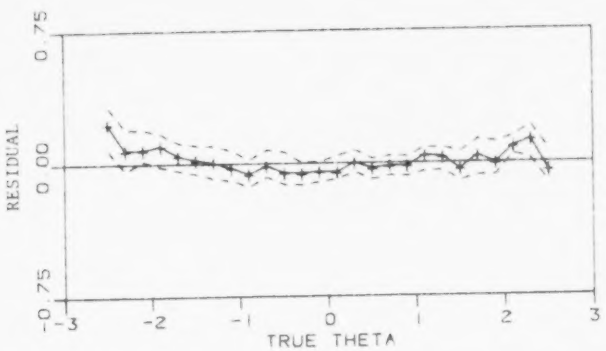
Figure 13: For cross-validation sample abilities estimated from true item parameters, the residual median estimated ability (solid curve) and the 5% two-tailed confidence interval (dashed curves) for the three more realistic tests. Typical test T1 is at the top; discriminating test T2 is in the middle; poorly discriminating test T3 is at the bottom.



Typical Test T1



Discriminating Test T2



Poorly Discriminating Test T3

Figure 14: For cross-validation sample abilities estimated from estimated item parameters, the residual median estimated ability (solid curve) and the 5% two-tailed confidence interval (dashed curves) for the three more realistic tests.

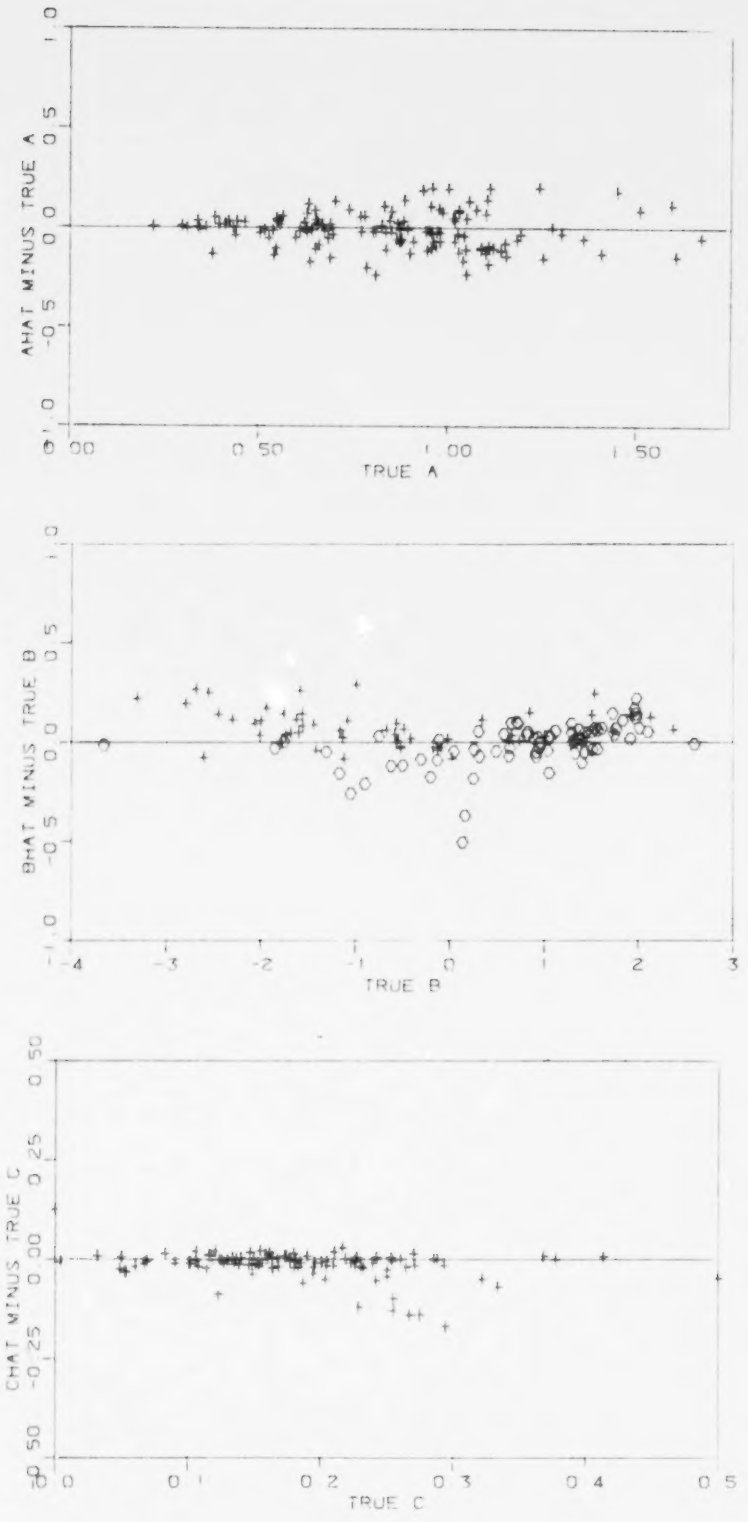


Figure 15: The residuals (estimated minus true) of the estimated item parameters for typical test T1.

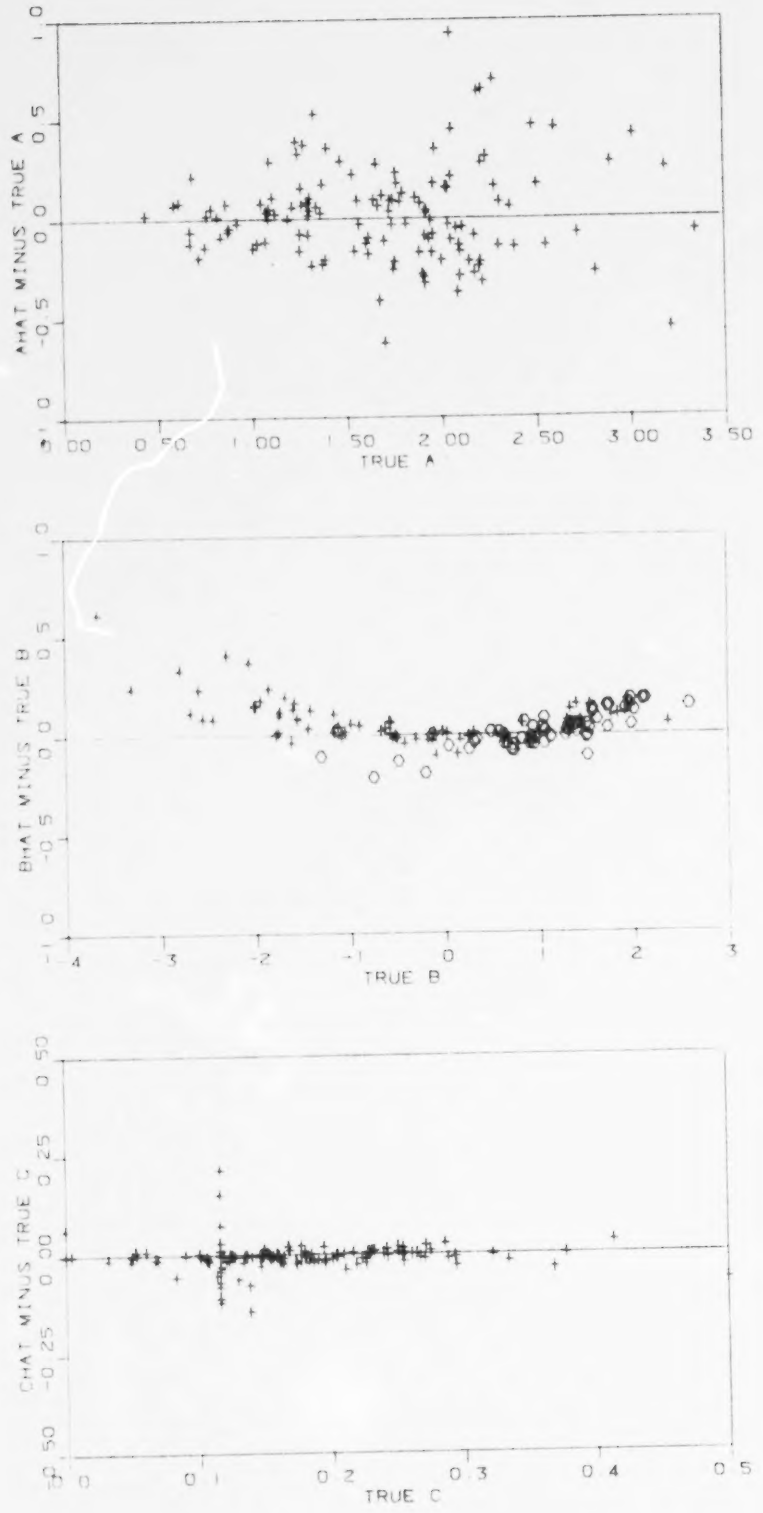


Figure 16: The residuals (estimated minus true) of the estimated item parameters for discriminating test T2.

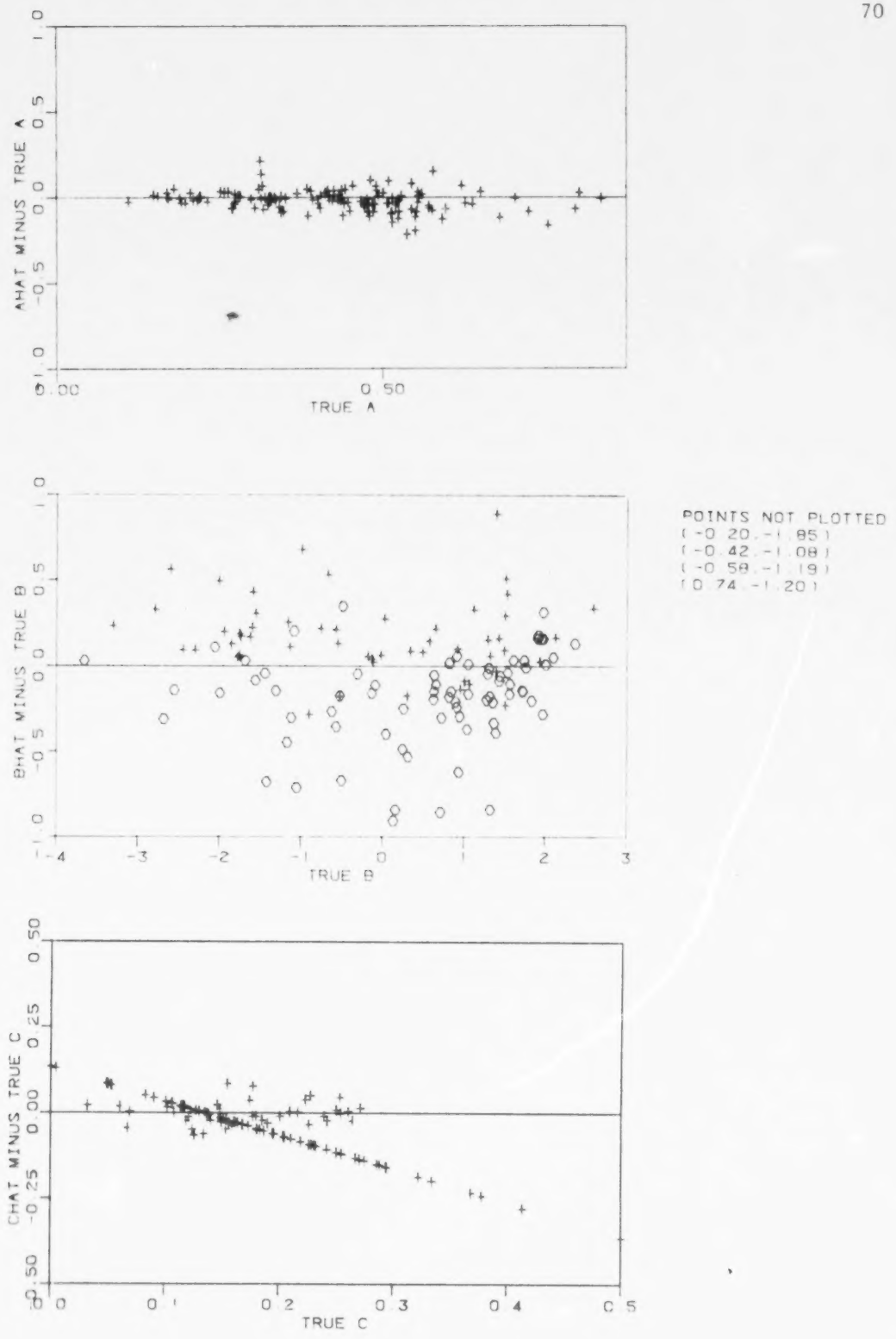
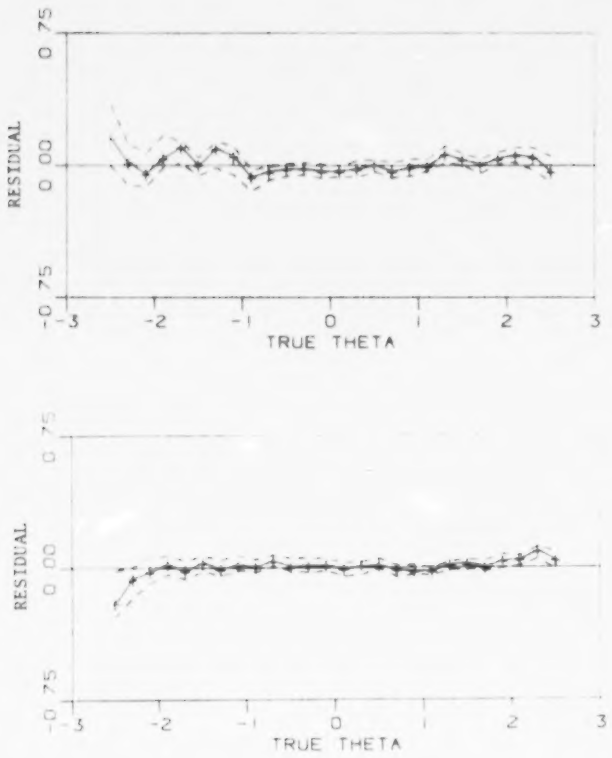


Figure 17: The residuals (estimated minus true) of the estimated item parameters for poorly discriminating test T3.



Typical Test T1

Discriminating Test T2

Figure 18: For calibration sample abilities estimated from estimated item parameters, the residual median estimated ability and the confidence interval for typical test T1 (top) and discriminating test T2 (bottom), when true a's are used as starting values.

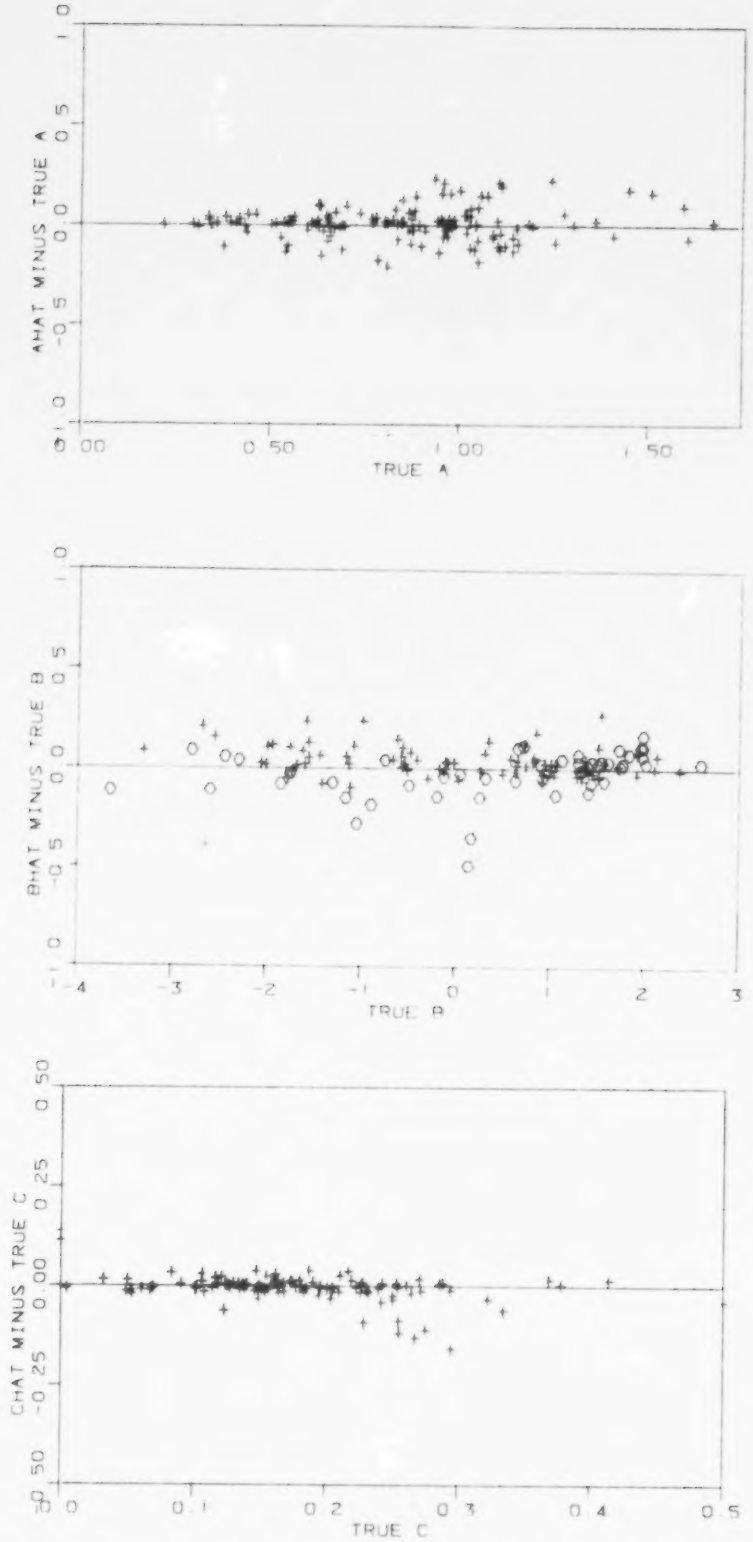


Figure 19: The residuals (estimated minus true) of the estimated item parameters for typical test T1 when true a's are used as starting values for the item discrimination estimates.

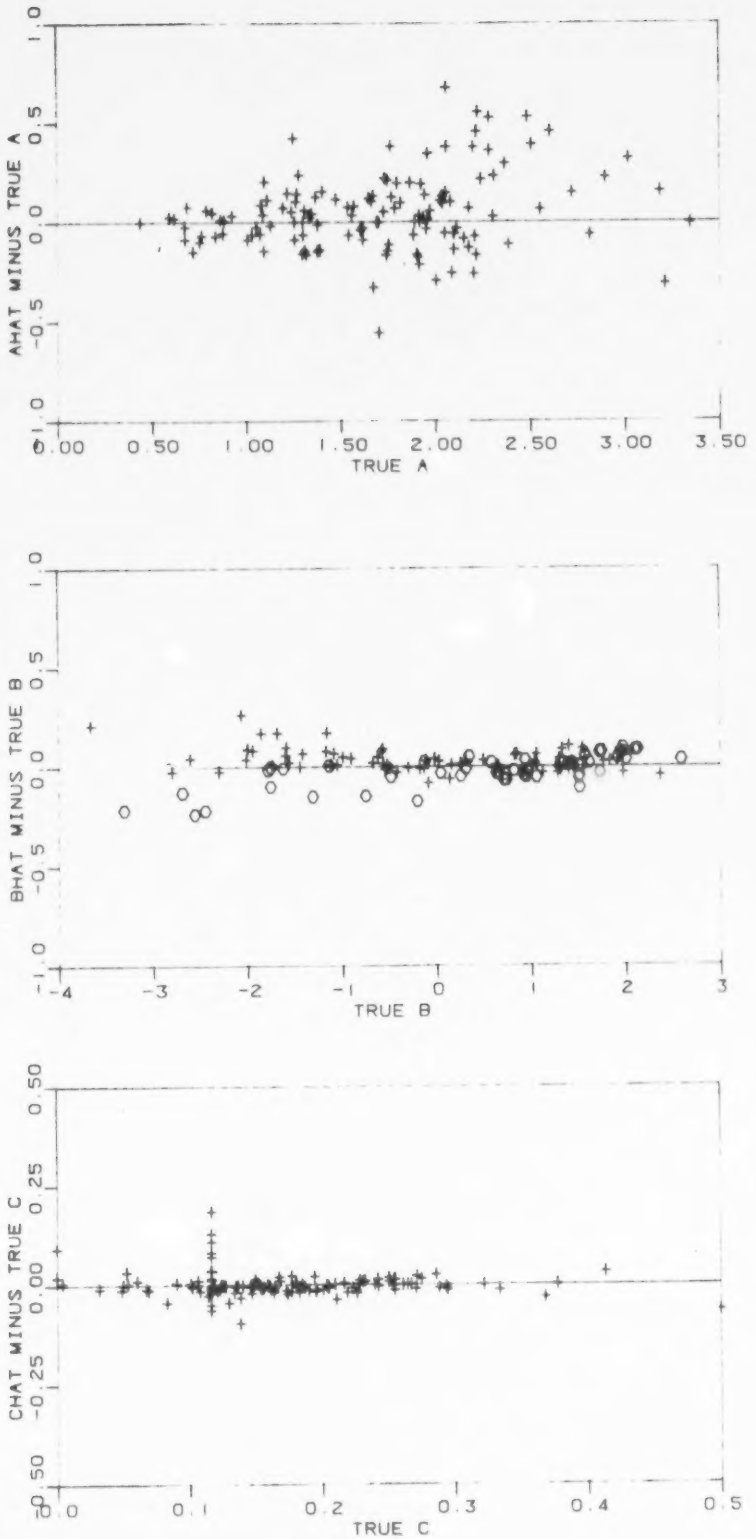


Figure 20: The residuals (estimated minus true) of the estimated item parameters for discriminating test T2 when true a's are used as starting values for the item discrimination estimates.

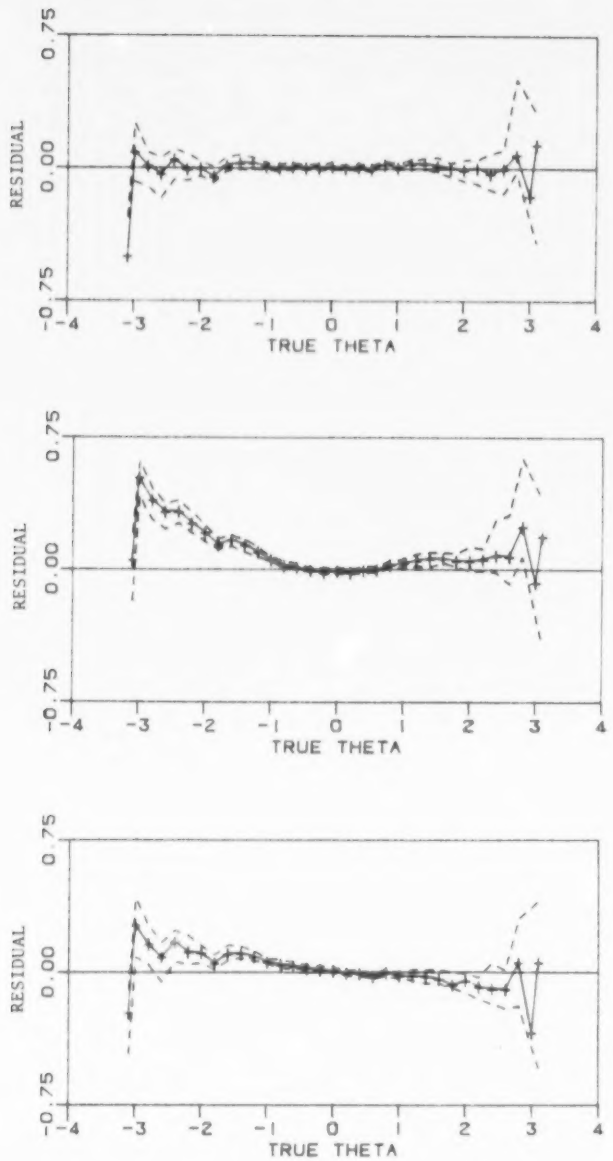


Figure 21: For ASVAB calibration sample abilities, the residual median estimated ability and the confidence interval when abilities are estimated from true item parameters (top), when abilities are estimated from estimated item parameters (middle), and when abilities are estimated from estimated item parameters in a calibration where true a's are used as starting values (bottom).

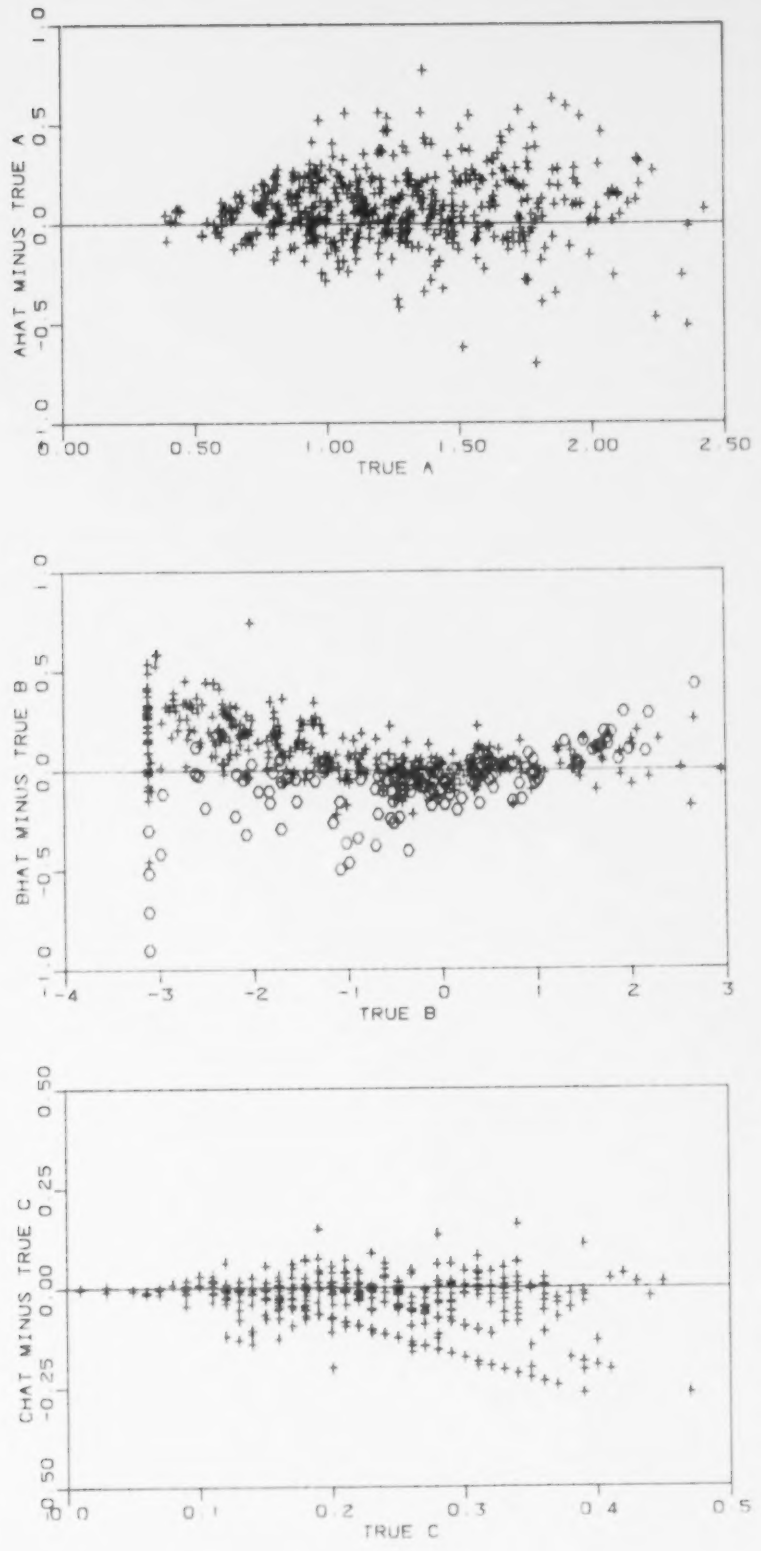


Figure 22: The residuals (estimated minus true) of the estimated item parameters for the ASVAB calibration.

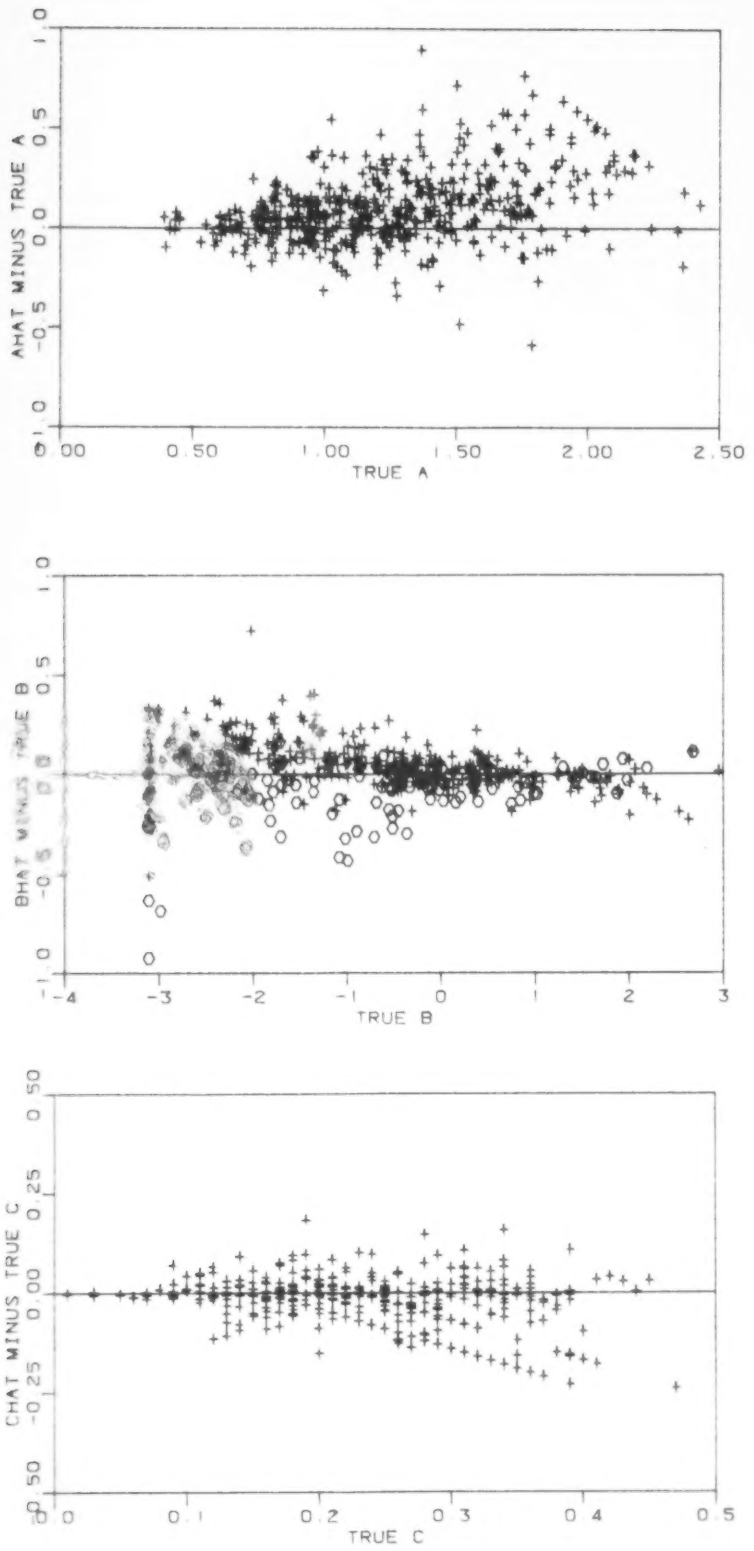


Figure 23: The residuals (estimated minus true) of the estimated item parameters for the ASVAB calibration when true a's are used as starting values for the item discrimination estimates.